THE FACULTY OF ENGINEERING AND THE BUILT ENVIRONMENT

POSTGRADUATE SCHOOL OF ENGINEERING MANAGEMENT

PREDICTIVE MAINTENANCE FRAMEWORK FOR CATHODIC PROTECTION SYSTEMS USING DATA ANALYTICS

by

Estella Barbara Rossouw

(200575264)

For a minor study in partial fulfilment for the degree

Magister Philosophiae

in

Engineering Management

At

THE UNIVERSITY OF JOHANNESBURG

Supervisor: Prof. Wesley Doorsamy

October 2020

ABSTRACT

Pipeline operators continuously seek to improve reliability, safety and reduce corrosion and maintenance costs. Cathodic protection (CP) systems is a secondary external corrosion-prevention mechanism for underground pipelines. Two types of CP exist, namely, sacrificial anode CP (SACP) or impressed current CP (ICCP). ICCP units consist of a rectifier that drives a current through an anode bed, to prevent the corrosion of the pipeline.

A study by the National Association of Corrosion Engineers (NACE) found that maintenance cost can escalate rapidly due to CP equipment damage, replacement of pipeline sections due to a damaged coating or forced corrosion, and ineffective time-based maintenance. This study presents and evaluates a predictive maintenance framework based on the conformance of the CP pipe potential to the NACE SP0169-2013 CP criteria for steel pipelines. The outcome of this study aims to reduce the maintenance cost, improve pipeline integrity and prevent catastrophic failures due to a pipeline rupture.

An empirical research methodology is selected for this study that utilizes historical data from remote CP stations for predictive modelling. The research context consists of two distinct pipeline sections, namely, a pipeline with downstream test posts and a Transformer Rectifier Unit (TRU), and a pipeline adjacent to a DC transit system rail with a Forced Drainage Unit (FDU). To understand the data and prepare it for analysis, activities such as data cleaning, feature engineering and data exploration is necessary before the predictive modelling evaluation.

The employed data analytics framework consists of machine learning and descriptive statistics to inform the research results. The CP pipe potential prediction with a multiple linear regression approach resulted in an RMSE of 0.153 for a TRU and 0.675 for an FDU with stray current. A classification model was developed using state labels for a CP pipe potential operating window and improved the RMSE (best result was 93.66% for an FDU with stray current). The downstream test post state was estimated by determining the current coefficients for the supplying ICCP unit and estimating the CP pipe potential based on the multiple linear regression formula for the ICCP unit. The pipeline health was estimated with ICCP and test post data, and the results were not linear but presented an overall error of 3%.

A maintenance matrix, consisting of the fault condition, risk and allowable time windows, was developed for the maintenance suggestion. A classification machine learning model predicted the required maintenance activity based on the state labels with the lowest accuracy of 96.64% (FDU), and highest of 99.67% (TRU), however, the time element was not considered. Time evaluation of suggested maintenance activities includes the Kaplan-Meier survival analysis, allowable cycle time analysis and time-series trend component analysis to forecast long-term maintenance requirements. The overall prediction results and suggested maintenance framework can potentially aid to reduce the maintenance cost of pipeline ICCP systems.

ACKNOWLEDGEMENTS

I would like to thank the following people for their support, prayers and motivation during the duration of this study:

i. My family, for always asking how things are going and supporting me throughout this process. Thank you for all your kind words of encouragement.
ii. My best friend and partner, thank you for all the love, support, massages and emotional guidance throughout this process.
iii. My supervisor, for providing technical guidance and going the extra mile to assist me.
iv. The Lord, for blessing me to study at a world-class University.
v. All my friends and colleagues for their loyal support.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

TABLE OF FORMULAS

Table of Formulas	
Formula Name	**Formula**
Metal Voltage (Gibbs)	$E = -\Delta GnF$
Adjusted DC Potential for CSE	$PA = PRE - PCSE + PReading\ vs\ RE$
Ohm's Law for Electrical Conduction	$V = \mathbf{IR}$
Electrical Resistivity (ohm)	$\mathbf{R} = \frac{\rho \ell}{\mathbf{A}}$
Mass Loss Rate (g/m2d)	$MR = K_2 \times I_{Corr} \times EW$

Table of Formulas	
Formula Name	**Formula**
Corrosion Rate (mol/m2-s)	$\mathbf{r} = \frac{i}{n\mathfrak{F}}$
Corrosion Penetration Rate(mm/yr)	$CPR = \frac{KW}{pAt}$
Risk Level Determination	$R = P \times C$
Temperature Compensation	$E_t = E_{25°C} + k_t \times (\mathrm{T} - 25°\mathrm{C})$
Current Density (mA/m2)	$\mathbf{i} = \sigma_e e$
AC Current Density (AAC/m2)	$\mathbf{i}_{ac} = \frac{(8V_{ac})}{\rho\pi d}$
Linear Regression Model	$\mathbf{y} = \beta_0 + \beta_1 x + e$
Multiple Linear Regression Model	$\mathbf{y}_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \cdots + b_p x_{ip} + e_i$
Youden's J Index	$J = Sensitivy + Specificity - 1$
CP Unit Risk Indicator	$R_{CPU} = ((U_T + R_U) - \frac{U_T + R_U}{U_O} + (F_{OP} \times R_{OP}) + (F_{UP} \times R_{UP}) + (F_P \times R_P)) \times \left(1 + \frac{D_U}{D_T}\right)$
CP Unit Risk Indicator – Inverse Effect	$R_{CPU} = ((U_T + R_U) - \frac{U_T + R_U}{U_O} + (F_{OP} \times R_{OP}) + (F_{UP} \times R_{UP}) + (F_P \times R_P)) \times \left(\frac{1}{1 + \frac{D_U}{D_T}}\right)$
CP Unit Health Indicator	$H_{CPU} = \left(1 - \frac{R_{CPU}}{T_{CPU}}\right) \times 100\%$
Overall Pipeline Section CP Health Indicator	$H_{CPO} = \frac{H_{CPU1} + H_{CPU2} + + H_{CPUn}}{n}$
Root Mean Squared Error	$RMSE = \sqrt{\frac{1}{n}\sum_{i=0}^{n} e_i^2}$
Mean Absolute Error	$MAE = \frac{1}{n}\sum_{i=0}^{n} \lvert e_i \rvert$
Survival Function	$S(t_j) = S(t_{j-1})\left(1 - \frac{d_j}{n_j}\right)$

TABLE OF ACRONYMS

Table of Acronyms	
Acronym	**Definition**
2-TBN	2-Step Temporal Bayesian Model
AC	Alternating Current
ACM	AC Mitigation Station
ADC	Analog-to-Digital Converter
AI	Artificial Intelligence
AMA	Autoregressive Moving Average
ANN	Artificial Neural Network
ANSI	American National Standards Institute
BstLm	Boosted Linear Model
CAPEX	Capital Expenditure
CARET	Classification And Regression Training
CBM	Condition-based Maintenance
CCEB	Current Condition Evaluation-Based
CFR	Code Of Federal Regulations
CIPS	Close Interval Potential Survey
CLI	Command-Line Input
CM	Condition Monitoring
CMS	Corrosion Management System
COM	Common
Corrplot	Correlation Plot
CP	Cathodic Protection
CPT	Conditional Probability Table
CRAN	The Comprehensive R Archive Network
CRISP-DM	Cross-Industry Standard Process For Data Mining
CSE	Copper-Copper Sulfate Electrode
CSV	Comma-Separated Value
DB	Database
DC	Direct Current
DCVG	Direct Current Voltage Gradient
ECDA	External Corrosion Direct Assessment
EMF	Electro-Motive Force
EMI	Electromagnetic Interference
ETTF	Estimated Time To Failure
FCPB	Future Condition Prediction-Based
FDU	Forced Drainage Unit
FMEA	Failure Modes and Effects Analysis
FTA	Fault-tree Analysis
FV	Future Value
gam	Generalized Additive Model
GDP	Gross Domestic Product
GIS	Geographical Information System
glm	Generalized Linear Model

Table of Acronyms	
Acronym	**Definition**
glmboost	Boosted General Linear Model
GPRS	General Packet Radio Services
GPS	Global Positioning System
GSM	Global System For Mobile Communications
GUI	Graphical User Interface
H-MOIA	Hybrid Multi-Objective Immune Algorithm
ICCP	Impressed Current Cathodic Protection
IDE	Integrated Development Environment
IJ	Insulated Joints
IoT	Internet Of Things
ISO	International Organization for Standardization
KDD	Knowledge Discovery In Databases
KM	Kaplan-Meier
k-NN	k-Nearest Neighbours
KPI	Key Performance Indicator
LAD	Logical Analysis Of Data
LFJ	Least Flexible Job First
LL	Long Life
lm	Linear Model
LPT	Longest Processing Time
LR	Linear Regression
LRR	Logistic Regression
LWSN	Linear Wireless Sensor Networks
MA	Moving Average
MAE	Mean Absolute Error
MCCV	Monte Carlo Cross-Validation
MIDOS	Minimal Impact Of The Disrupted Operation On The Schedule
MILP	Mixed Integer Linear Programming
ML	Machine Learning
MLP	Multilayer Perceptron
MOV	Metal-Oxide Varactor
MRG	Methane Rich Gas
MRUL	Mean Remaining Useful Life
MTBF	Mean Time Between Failures
MTTF	Mean Time To Failure
MTTR	Mean Time To Restore
MTU	Master Terminal Unit
N/A	Not Applicable
NA	Not Available
NACE	NACE International, Formerly National Association Of Corrosion Engineers
NDU	Natural Drainage Unit
NPV	Net-Present-Value
OLE-DB	Object Linking And Embedding, Database
OP	Over-protected

Table of Acronyms	
Acronym	**Definition**
OPEX	Operational Expenditure
OSA-CBM	Open System Architecture For Condition-Based Maintenance
OW	Operating Window
P	Protected
PB	Protection Band
PCM	Pipeline Current Mapper
PCS	Process Control Systems
pdf	Probability Density Function
PdM	Predictive Maintenance
PHM	Proportional Hazards Model
PIMS	Pipeline Integrity Management System
PIR	Period Inspection And Replacement
PM	Preventative Maintenance
POF	Physics of Failure
PV	Process Value
QIR	Quantile Based Inspection And Replacement
RBI	Risk-based Inspection
RBM	Risk-based Maintenance
RCM	Reliability-Centred Maintenance
RE	Reference Electrode
RF	Random Forest
rlm	Robust Linear Model
RMSE	Root Mean Squared Error
ROI	Return-On-Investment
RTU	Remote Terminal Unit
RUL	Remaining Useful Life
RVM	Relevant Vector Machines
SACP	Sacrificial Anode CP
SAS	Standard Analytical Software
SCADA	Supervisory Control And Data Acquisition
SCC	Stray Current Corrosion
SCE	Calomel Reference Electrode
SCR	Silicon-Controlled Rectifier
SHE	Standard Hydrogen Electrode
SL	Short Life
SOLO	Structure Of Observed Learning Outcome
SP	Setpoint
SPC	Statistical Process Control
SPSS	Statistical Package For The Social Sciences
SQL	Structured Query Language
SSC	Silver-Silver Chloride Reference Electrode
SVM	Support Vector Machine
svmLinear	Linear Support Vector Machines
SVR	Support Vector Regression

Table of Acronyms	
Acronym	**Definition**
TBM	Time-based Maintenance
TP	Testpost
TRU	Transformer Rectifier Unit
UP	Under-protected
ZRE	Zinc Reference Electrode

TABLE OF STANDARDS CONSULTED

Table of Standards Consulted	
Standard	**Description**
ASTM G96-90	Standard Guide For Online Monitoring Of Corrosion In-Plant Equipment (Electrical And Electrochemical Methods)
NACE SP0169-2013	Control Of External Corrosion On Underground Or Submerged Metallic Piping Systems
NACE TM0497-2018	Measurement Techniques Related To Criteria For Cathodic Protection On Underground Or Submerged Metallic Piping Systems
CFR PART 192	Transportation Of Natural And Other Gas By Pipeline: Minimum Federal Safety Standards
RP0104-2004	The Use Of Coupons For Cathodic Protection Monitoring Applications
PAS 55-1:2008	Asset Management
CSA Z662-2007	Safety Standard - Oil And Gas Pipeline Systems
ASME B31.8	Gas Transmission And Distribution Piping Systems
ISO 31000	International Standard For Risk Management
ISO 17359:2018-01	Condition Monitoring And Diagnostics Of Machines — General guidelines
ISO 13372	Condition Monitoring And Diagnostics Of Machines — Vocabulary
ISO 13374-1	Condition Monitoring And Diagnostics Of Machines — Data Processing, Communication And Presentation — Part 1: General Guidelines
ISO 13374-2	Condition Monitoring And Diagnostics Of Machines — Data Processing, Communication And Presentation — Part 2: Data Processing
ISO 13374-3	Condition Monitoring And Diagnostics Of Machines — Data Processing, Communication And Presentation — Part 3: Communication
ISO 13374-4	Condition Monitoring And Diagnostics Of Machines — Data Processing, Communication And Presentation — Part 4: Data Presentation
ISO 13379-1	Condition Monitoring And Diagnostics Of Machines — Data Interpretation And Diagnostics Techniques — Part 1: General Guidelines
ISO 13379-2	Condition Monitoring And Diagnostics Of Machines — Data Interpretation And Diagnostics Techniques — Part 2: Data-Driven Applications
ISO 13381-1	Condition Monitoring And Diagnostics Of Machines — Prognostics — Part 1: General Guidelines

CHAPTER 1: INTRODUCTION AND BACKGROUND

1.1. Introduction

Underground pipelines are critical infrastructure in an economy to provide piped products such as water, oil, and gas to industrial and residential consumers. In most urban areas, a variety of different product pipelines spans across densely populated areas [1]. Pipeline networks extend past the perimeter of the various metropolitan regions to provide an interconnect between the product source and the consumer. Product sources consist of storage facilities, gas extraction plants, water and sanitation networks, and petrochemical industries. These pipeline networks often cross various public and private industrial, residential, and agriculture property to transfer the product between the source and destination [2].

According to The Charleston Advisor, South Africa had a total installed underground pipeline network of 3839km in 2013 [3]. Future expansion of industries will require an ever-increasing underground pipeline network [2]. Albeit the development of underground pipeline networks as a stimulus to economic activity, corrosion poses a significant risk that can decrease the asset's lifetime, result in substantial product loss due to accidents, failures, or loss of production, and most significant result in loss of life [4].

Corrosion is a natural phenomenon where a material deteriorates due to its interaction with the environment and is inevitable due to a material's fundamental need to reduce its energy state to a lower oxide-state (Gibbs Energy) [5]. The National Association of Corrosion Engineers (NACE) estimated in 2013 that the cost of corrosion was US2.5$ trillion, which equated to 3.4% of the global Gross Domestic Product (GDP) [4].

Implementation of a Corrosion Management System (CMS), based on the ISO 31000 risk-management standard, can provide an estimated 15%-35% savings on the cost of corrosion. Cost savings include a reduction in maintenance and inspection costs, a decrease in pipeline ruptures resulting in less product loss, and an extension of the asset's life, which delays capital expenditure to replace pipeline sections. The CMS should include the corrosion threat in all phases of the asset's life cycle, as the cost of unchecked corrosion can have a significant impact on an organization when the asset is in use. Apart from the financial, environmental, and public safety impact of corrosion, pipeline operators can lose their operating license, which could have a negative impact on economic activity [6].

NACE defines three methods for mitigating corrosion, which includes a change in the environment, a change in material, or placing a barrier between the material and the environment [7]. The barrier consists of a pipeline coating or wrapping, which prevents contact between the pipeline and the electrolyte [2],[4]. The area around the pipe can also be compacted with backfill material to prevent a collapse of the pipe trench [8],[9].

Cathodic Protection (CP) is a secondary corrosion control mechanism that uses direct current (DC) to control external corrosion of underground pipelines. Rectifiers installed along an underground pipeline provides the required DC for effective corrosion control.

The CP system forces all anodic areas of the pipe to cathodic areas by impressing DC onto a dedicated anode, which will corrode instead. This process is referred to as "polarization" and results in a quasi-equilibrium potential difference condition between regions [10]. If all pipeline regions are cathodic, no corrosion will occur [7]. The CP system forms part of the CMS to reduce the pipeline corrosion risk [10].

Corrosion monitoring activities such as inspections and monitoring are vital to establish the pipeline integrity state and manage the associated risk. In-line and on-line data collection enables state determination and pre-empts proactive management of the corrosion risk [5]. The NACE standard, SP0169-2013, provides the CP pipe potential criteria for operating a pipeline and includes the required maintenance activities and pipeline monitoring guidelines [11].

Operating extensive pipeline networks can be achieved through the implementation of a Supervisory Control and Data Acquisition (SCADA) system, which monitor and control remote sensors over a large geographic area through data acquisition, networked data communication, data presentation, and control [12]. Analysis of stored data can highlight process inefficiencies and allow for optimization [13].

Pipeline monitoring can include a variety of sensors to monitor impressed-current CP (ICCP) rectifiers remotely and typically only include periodic measurement of pipe-to-soil potential between rectifiers to determine the effectiveness of the CP system [11]. Most CP maintenance strategies utilized by the industry currently focus on time-based preventative maintenance through periodic visual inspection and scheduled recordings of CP pipe potentials as recommended by the SP0169-2013 standard. Reactive maintenance (run-to-failure) concentrates on replacing damaged CP equipment.

Data analytics is defined as the process to analyze large data sets to support decision making. A simplified CRISP-DM approach allows for knowledge extraction from databases through data preparation, pre-processing, analysis, and post-processing [13]. Implementation of data analytics can assist in predicting the future operating state of a piece of equipment based on its historical data [14]. SCADA systems usually provide a software driver to store data into a relational database.

This study aims to use data analytics to evaluate the prediction capability of required maintenance activities and the state of ICCP units and downstream test posts (TP) by using only the CP pipe potential and the rectifier output measurements as predictors. The data for this study consists of historical SCADA and logger data from both CP rectifiers and TP's. The output of this research is to bring forth a predictive maintenance framework to be incorporated into a maintenance strategy or CMS with the desired effect to reduce OPEX costs for maintaining CP systems.

1.2. Problem Statement

Based on the NACE Economics of Corrosion study, the OPEX cost to maintain underground pipelines increases rapidly over time if a proper CMS system is not in place or if the CMS was not part of the original asset design [6]. The increase in OPEX cost relates to increased maintenance requirements, CP equipment damage,

replacement of pipeline sections, product loss, adverse environmental impact due to product spillage and fatalities or hospitalization of personnel and the public [6].

The cost to perform CP maintenance on a planned schedule for significant pipeline networks is not sustainable in the long run, and the use of new technological advances in data analytics requires investigation to promote sustainable and safe pipeline operations.

1.3. Research Aim

This study aims to evaluate the feasibility of predicting the ICCP unit and TP state and maintenance suggestions for existing pipeline CP systems through a combination of reliability management principles, risk management, and historical data analytics. Data analytics of existing pipeline CP systems, where design information might be missing, is significant since the results can assist in decision-making primarily through the past operation of a specific ICCP unit.

Data to be used for the analytics will consist of CP data from SCADA systems, manual recordings, and asset information. The resultant output should predict the required maintenance and state of the ICCP unit and downstream TP's based on specific operating conditions and the associated risk.

The predictive state and maintenance results from this study, should motivate its inclusion in a maintenance strategy or CMS, and reduce OPEX costs, improve pipeline integrity, and promote sustainable pipeline operations. This study will focus primarily on underground pipeline networks with CP installed.

1.4. Research Objectives

The primary aim of this study is to establish and evaluate a predictive maintenance framework for pipeline ICCP systems, based on the NACE SP0169-2013 standard and using only historical CP operating data. The presented predictive maintenance framework should consider both the ICCP unit state and related maintenance activity.

To achieve the research aim, the following research objectives are defined:

1. Determine if statistical analysis of CP data, based on the NACE SP0169-2013 criteria for CP evaluation, can predict or estimate the ICCP unit and downstream TP state.
2. Determine if a maintenance activity can be suggested based on the ICCP unit state.

1.5. Research Questions

To achieve the objectives and to provide guidance for this study, the following research questions are defined:

1. Which statistical analysis methods can be used on historical and real-time CP data to predict or estimate the state of ICCP units or TP's?
2. Which maintenance activities are required to remedy the ICCP unit state, and what mechanism can be used for suggesting maintenance activities?

1.6. Scope Of The Study

The primary focus areas of this study will be limited to underground pipeline networks and include the following:

- Review the literature related corrosion theory, external corrosion prevention of underground pipelines, standards and statutes governing pipeline operations, CP system operation and maintenance strategies, reliability engineering principles, maintenance strategies, data analytics approaches and case studies related to the scope of this study.
- Evaluate the ICCP unit data for different operating conditions.
- Evaluate the ICCP and downstream TP states based on conformance to a pre-defined OW.
- Inspect the data for this study, add additional features, remove erroneous columns and rows, and perform an exploratory data analysis.
- Design, test and evaluate the descriptive statistics based on the selected pipeline section.
- Perform time-series analysis on the CP pipe potential, evaluate the trend component and assess its use for maintenance suggestion.
- Perform multiple linear regression analysis on the ICCP unit data, evaluate the prediction accuracy of the CP pipe potential, and select the model with the best accuracy.
- Evaluate a classification model in R to predict the CP pipe potential state and evaluate the prediction accuracy.
- Develop a maintenance matrix that considers the defined ICCP unit states, risk factor and time limit.
- Evaluate the prediction accuracy of the maintenance activity suggested and perform time analysis using the Kaplan-Meier curve, cycle time and time-series analysis.
- Critically evaluate the results and the proposed predictive maintenance framework and summarize findings.
- Provide recommendations and suggest future work.

1.7. Methodology

1.7.1. Research Approach And Design

The research approach and design consist of the following:

i. An empirical research design that utilizes historical CP data for predictive modelling.
ii. A research context that considers two unique pipeline sections to evaluate the prediction accuracy for different ICCP units and downstream TP's.
iii. Data exploration and feature engineering activities to ensure the data is in the correct format to perform the analysis.
iv. Evaluation of the study's research objectives through the development and testing of machine-learning (ML) models using R Studio.
v. ML model performance evaluation to conclude on the feasibility of this study's predictive modelling objectives. The prediction accuracy was benchmarked

against pre-defined performance metrics, such as the RMSE, MAE and the percentage error to determine if the models provide accurate prediction results.

1.7.2. Data Collection

Data for this study consisted of historical CP operating data from both a SCADA system and manual recordings database.

1.7.3. Limitations

The data utilized in the study consisted only of CP pipe potentials and the rectifier output current, output voltage and drainage current for a natural gas pipeline in South Africa. The CP pipe potentials differ in magnitude in South Africa when compared to countries such as the United States of America (due to the DC transit systems used in South Africa). The limitation of the data for this study is the defined OW (which requires adjustment based on unique operating conditions).

The presented predictive maintenance framework only evaluated nine machine-learning techniques for linear regression and three classification techniques to predict the CP pipe potential, the CP pipe potential state and the suggested maintenance activity. Different machine-learning techniques exist that can improve the prediction accuracy and reduce computational load (as evident in the list of available techniques in the Caret package [15] and tabled in Appendix D2).

The data-driven approach of the presented framework did not consider as-built CP system design information, as only historical CP operating data was available. Since the scope of this study is limited to CP systems (external corrosion control), internal corrosion monitoring and modelling were not included in this study. The study did also not consider the IR-Drop and Temperature Effects of the CP potential (no data was available).

1.7.4. Ethical Considerations

No ethical considerations.

1.8. Significance Of Research

There is a significant drive from all industries to become more cost-effective to survive in a challenging global economy. For pipeline operators with extensive pipeline networks (> 1000 km), operational costs are high due to the large geographical area in which the pipeline network runs. Cost drivers include the significant geographic area of the pipeline network and the associated maintenance requirements, poor system design, and an inherently unreliable system. The complexity of CP systems is also a key contributor to high maintenance costs.

The proposed research will enable the statistical prediction of an ICCP unit or TP state and maintenance required based solely on historical and real-time operating CP data. This prediction capability should reduce maintenance and operational costs by streamlining maintenance activities of CP systems.

1.9. Role Of The Candidate

The role of the candidate is as follows:

- Plan the research timelines
- Collect the required data for the statistical analysis
- Perform the quantitative study
- Compare results and conclude on study outcome

1.10. Research Assumptions

The assumptions for the research is as follows:

- The data collected is accurate and is indicative of a fully-functional CP system.
- Any simulated data will be based on an empirical study or estimated from the CP system operational state.
- Although the research applies to underground gas pipelines, the related implementation should be similar for other pipelines with identical design characteristics

1.11. Research Layout

The research layout is as follows:

Chapter 1: Introduction

Chapter one provides an introduction to the research, the problem statement, the aim, and significance of the research. Furthermore, this chapter also outlines the research methodology, assumptions and ethical considerations.

Chapter 2: Literature Review

Chapter two consists of a detailed literature review which investigates the principles of corrosion and corrosion prevention. The latter includes CP systems, which evaluates the operation and maintenance according to various standards and regulatory requirements. A review of a CMS framework and pipeline integrity management system (PIMS) principles seeks to determine if the predictive modelling and maintenance approach will fit into existing CMS or PIMS frameworks.

Reviewing reliability engineering theory seeks to determine potential overlaps for this study in terms of data analysis, condition monitoring and maintenance strategies. Furthermore, this chapter also investigates the use of various data analytics techniques applicable to predictive modelling and maintenance. Lastly, a review of case studies was required to determine possible research design methodologies applicable to this study.

The literature review consists mainly of theoretical concepts based on the limited research available for the scope of this study.

Chapter 3: Research Design and Methodology

Based on the literature review, this chapter describes the research methodology and design approach for this study and elaborates on the research context and data

collection methods. The data analysis section is key to this study and describes various activities such as data cleaning, feature engineering, prediction and learning and performance evaluation. The section concludes with a review of the study's limitations and delimitations.

Chapter 4: Exploratory Data Analysis

An exploratory data analysis is performed in chapter four and seeks to identify more information on the operation of ICCP units under different operating conditions. Furthermore, this chapter also evaluates the feasibility of a CP health indicator for a pipeline section based on ICCP operation and downstream TP's (as designed in chapter three).

Chapter 5: Predictive Modelling Evaluation and Results

Following the results from chapter four, this chapter evaluates the predictive modelling of CP pipe potentials, ICCP unit state, time-to-state analysis and maintenance suggestion dependant on specific operating conditions.

Chapter 6: Conclusion and Future Work

Chapter six synthesizes the results with the research objectives and questions and conclude with recommendations and future work.

1.12. Conclusion

This chapter discussed the basic outline for this study and reviewed the problem statement, research design and research objectives and questions.

Motaghare et al. suggest that predictive maintenance can streamline maintenance requirements and also reduce system downtime [14]. The candidate aims to present a predictive maintenance framework for ICCP units and TP's to reduce the run-to-failure or time-based maintenance costs associated with significant underground pipeline networks.

The next chapter reviews the literature applicable to this study, as applicable to the related research questions.

2. CHAPTER 2: LITERATURE REVIEW

2.1. Introduction

The literature review is a formative review of academic literature related to the field of study. The literature review should aim to be informative, provide an unbiased synopsis to the reader, and provide a balanced view of the related literature available for the study (which also indicates inconsistencies). The literature review is a pivotal step in forming a research idea, identifying areas for improvements, and determining the study's possible contribution [16].

According to Snyder, the literature review process consists of four stages, namely the designing the review, conducting the review, analysis and synthesis and lastly writing the literature review [17]. The summary presented below describes the key focus areas of the literature review for this study.

Due to the limited information available for predictive maintenance of CP systems, the literature review consists of the following sections:

i. Review the theory applicable to corrosion, such as the basic corrosion cell, electrode potentials, corrosion types, and the fundamental aspects of corrosion prevention.
ii. Building on the corrosion theory literature, the prevention, monitoring, and measurement of external corrosion using CP systems was deemed necessary and taking into account the applicable industry and statutory standards. This section also explores the use of remote monitoring solutions for CP monitoring.
iii. The next section briefly investigates the use of integrity management systems for pipelines and provides a short overview of a typical system's composition.
iv. A recent study by NACE presented findings on the use of an integrated CMS and its role in reducing the corrosion costs. This section reviewed the typical CMS building blocks.
v. The next section evaluates the reliability engineering principles relevant to the study and investigates different maintenance strategies applicable to the study's scope.
vi. The data analytics section delves deeper into the aspect as to how a predictive maintenance system can be developed by investigating different statistical learning approaches.
vii. The last section reviews several case study's relevant to predictive maintenance design and software implementation in other industries (none were available for CP systems).
viii. This chapter concludes with a summary of the literature review findings and the motivation for the study.

2.2. Corrosion Theory

This section evaluates the fundamentals of corrosion and external corrosion control techniques.

2.2.1. Corrosion Basics

Corrosion is a natural phenomenon whereby a metal corrodes due to interaction with its environment [18]. The "environment," typically referred to as the electrolyte usually consists of soil for underground pipelines. All metals are prone to degradation over time, and various corrosion types exist (dependant on the use and environment) [19].

According to Peabody, the process of corrosion relates to the field of thermodynamics. Metals extracted in ores require a significant amount of energy in the extraction process, which places metals in a high-energy state (according to the Gibbs Free Energy Theory [5]). Ores are typically oxides of extracted metals. The thermodynamics principle states that these metals are in a highly unstable state and will seek to achieve a lower energy state (such as an oxide or other compound). This process of seeking a lower energy state or oxidation is also known as corrosion [18].

The corrosion process consists of two electrochemical reactions, namely oxidation (loss of electrons) at the anodic site and reduction (consumption of electrons) at the cathodic site. When oxidation occurs, a negative charge develops between the electrolyte and the metal. The reduction reaction neutralizes this negative charge. The reduction reaction must take place to prevent a large negative charge and to cease the corrosion process [18].

Oxidation and reduction reactions are also referred to as "half-cell reactions" and can occur locally or separated by a physical distance. The physical separation of half-cell reactions is called a "differential corrosion cell" [18].

A direct electric current flows between the anode and the electrolyte at the anodic sites due to metal ions leaving the anodic sites. The flow of these ions to the place where reduction occurs is known as the cathodic site. According to Peabody, a corrosion cell only forms if all elements of the corrosion cell are present, namely, the anode, cathode, electrical conductive electrolyte, and a metallic connection between the anode and cathode [18].

The NACE practical galvanic series suggests that the cathode is usually at a higher potential than the anode, and conventional current flow occurs due to electron flow from the anode to the cathode [20].

Figure 2-1illustrates the composition of an elemental corrosion cell:

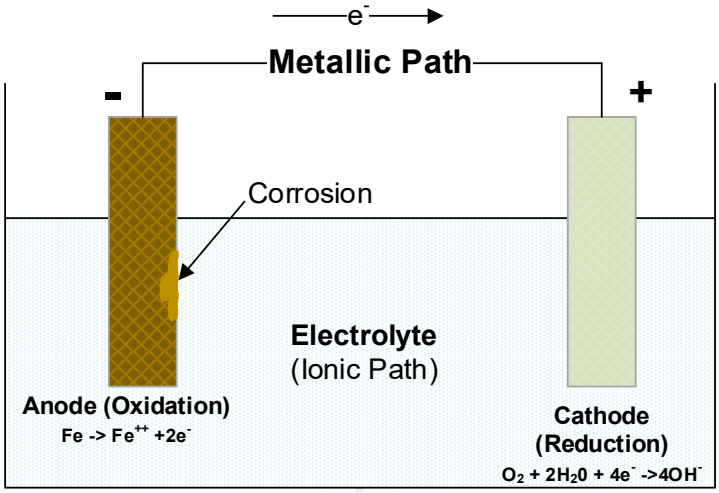

Figure 2-1 – Basic Corrosion Cell - Source: Adapted from [18]

Examples of reduction reactions include [21]:

$$H^+ \rightarrow \frac{1}{2}H_2 - e^-$$

$$Cu^{2+} \rightarrow Cu - 2e^-$$

$$Fe^{3+} \rightarrow Fe^{2+} - e^-$$

Examples of oxidation reactions include [21]:

$$Zn \rightarrow Zn^{2+} - 2e^-$$

$$Al \rightarrow Al^{3+} - 3e^-$$

$$Fe^{2+} \rightarrow Fe^{3+} - e^-$$

If all elements of the corrosion cell are present, a voltage will develop across the two half-cells. This potential difference or voltage is the driving force for the half-cell reactions. The magnitude of the potential is determined by the metal types in the corrosion cell and will differ for different pairs of metals [22].

The potential difference in the example below develops where the iron (Fe) metal is the anode (corrodes), and the copper (Cu) metal is the cathode (electrodeposits) [22].

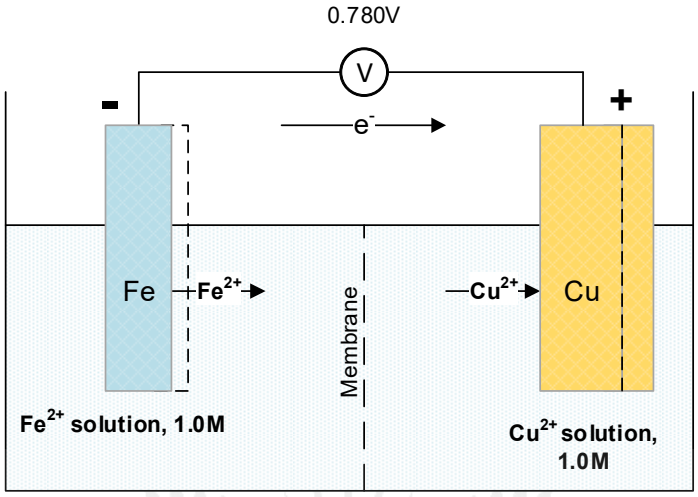

Figure 2-2 - Corrosion Cell with Potentials - Source: Adapted from [22]

In underground pipelines, corrosion usually occurs due to differential corrosion cells forming along the pipeline. Typical differential corrosion cells include pipeline exposure to mixed levels of oxygen concentration in the electrolyte (soil) or pipe-surface differences or varying soil chemistry [18].

2.2.2. Electrolyte

The electrolyte can consist of various compositions, such as soil (neutral, well-aerated, wet, heated, and acidic), seawater, and stationary and moving freshwater [23].

The electrolyte conductivity directly impacts the corrosion rate of a metal placed in a specific electrolyte. Electrolyte conductivity measurements facilitate estimating the metal corrosion rate in a particular electrolyte [23][22]. For further reading, the ASTM G96-90 standard can be consulted to evaluate the impact of the electrolyte conductivity on the metal corrosion rate [24].

2.2.3. Electrode Potentials

Callister refers to metals placed in a corrosion cell as "electrodes," and each metal has a unique native DC potential [22]. Inzelt et al. further expand on the definition of electrodes as either an electron conductor or the half-cell between an electron and ion conductor, respectively. Reference electrodes (RE) can reproduce a stable Galvani DC potential difference (either primary or secondary) [25].

This section discusses the standard/native DC potentials of electrodes and the DC potential when measured with reference to a reference electrode.

2.2.3.1. Electrode Standard/Native Potential

Yang suggests a metal's DC potential is related to the Gibbs Free Energy charge theory, which presents a formula for determining the DC potential of a specific metal [26]:

$$E = -\frac{\Delta G}{nF} \qquad \textit{2.1 Metal Voltage (Gibbs)}$$

Where:

- E = Potential in V_{DC}
- ΔG = Gibbs free energy charge of metal
- n = number of electrons transferred
- F = Faraday's constant

2.2.3.2. Standard Electromotive Force Series

The electromotive force (EMF) presents various definitions in the literature. Applicable to this study's scope, Inzelt et al. refer to the EMF as "*the limiting value of the electric potential difference of a galvanic cell when the current through the external circuit of the cell becomes zero*" [25].

According to Fowler and Lewicki, the corrosion rate of different metals in an electrolyte can be estimated based on the EMF of the two metals. Determining a metal's EMF (in a laboratory) requires a potential measurement with reference to a standard hydrogen reference electrode (SHE).

The EMF series consists of metals arranged in a series of descending values [20]. Metals with a higher DC potential value have an increased ability to release electrons. The higher the DC potential difference between the electrodes, the greater risk of corrosion exits due to increased current flow between the electrodes [22]. The EMF Series is tabled in Appendix A1.

The EMF series only considers pure metals, and in practice, most metals used are compounds or alloys. The galvanic series, however, considers metal compounds and is utilized in the industry [27].

2.2.3.3. Galvanic Series

In the galvanic series, a standard $CuCuSO_4$ reference electrode determines a metal's potential, while the system is in a state of equilibrium (no corrosion taking place). Similar to the EMF series, the arrangement of metals in the galvanic series depends on the metal being either active (will corrode) or noble (will not corrode) [27].

Fontana et al. suggest careful consideration of the selection of two metals for an application since the potential difference between two metals are the driving force for corrosion and should be limited. [27]. The galvanic and EMF series is thus significant in corrosion prevention system design (for metal-pair selection) [28].

Callister presented the galvanic series of various metal alloys used in the industry today and is presented in Appendix A2.

2.2.3.4. Reference Electrodes

As mentioned in the preceding section, a RE is a device that can reproduce stable Galvani DC potentials. RE's is a critical element to determine an electrode's DC potential and is significant to this study's scope (when determining the CP pipe potential).

Various literature sources suggest that the Standard Hydrogen Electrode (SHE) is the primary reference electrode and utilized to calibrate secondary electrodes. According to Park, various reference electrodes are available for different environments (for example, freshwater vs saltwater). Each electrode will provide a unique metal/interface DC potential based on its composition [29]. The SHE's features such as ease of preparation, ability to attain a fast equilibrium state, non-polarizable properties, and environmentally friendly composition makes it the fundamental reference electrode [30].

The primary reference electrode consists of a piece of metal immersed in a solution of one of its salts. Thermodynamically stable reference electrodes follow the Nernst Equation and will have a known reversible chemical reaction between the metal and its surrounding environment. At a state of equilibrium, the rate of chemical reactions in both directions will be equal [31].

Ansuini and Dimond suggest that the two most common reference electrodes consist of either metal in a solution of dissolved ions of the metal or a submerged metal coated with a metal's salt. The former being a copper/copper sulfate (Cu/CuSO4) electrode and the latter either a silver/silver chloride (Ag/AgCl) or calomel (mercury/mercury chloride) electrode [31].

Two categories of reference electrodes are available in the market, namely, portable and stationary reference electrodes. The former used for a specific time and the latter for extended periods [28].

2.2.3.5. Reference Potential

The DC potential of an electrode depends on the RE element metal and electrolyte composition [31]. The DC potential of reference electrodes (E) are expressed to the SHE and includes the sign convention (positive or negative). Only in cases where a standard electrode DC potential is studied at specific conditions (saturation of ion concentration and the temperature is 25 ºC), the electrode DC potential will be the standard electrode DC potential (E^0) [30].

Comparison of a RE DC potential scale for each reference electrode relative to a SHE is required when using another RE as different RE's are at different DC potentials. The reference electrode used must be noted when taking DC potential measurements to ensure that the DC potential offset is considered [31].

RE's have common characteristics that indicate whether the electrode is in a proper working condition. McCafferty suggests that the three main aspects are constant half-cell DC potentials; stable half-cell DC potential if current pass through the cell; half-cell DC potential should not drift over time [32]. NACE recommends a standard

maintenance procedure for CSE electrodes. Maintenance activities include cleaning, proper storage, periodic replacement of metal rods, and copper sulfate solution; ensure the copper sulfate solution is not contaminated [33].

The relative potentials of common RE's are tabled in Appendix A3 and conversion of potentials from one RE to another is tabled in Appendix A4.

2.2.3.6. Potential Measurements across a Metal/Electrolyte Interface

Determining the DC potential difference between a metal and an electrolyte/solution is possible by taking a potential measurement with reference to a standard RE or half-cell [27]. The governing equation for electrical conduction is Ohm's Law, which Callister suggests *"relates the current I—or time rate of charge passage—to the applied voltage V"* [34].

$$V = IR \qquad \textit{2.2 Ohm's Law for Electrical Conduction}$$

Where:

- V = Voltage in V_{DC}
- R = Resistance of material in Ω
- I = Current flow through the material in A

High input impedance (>10MΩ) voltmeters enable measurement of electrode potentials, and will only allow a minimal current to flow through the voltmeter's measurement circuit. The effect of this current through the measurement circuit is negligible. The voltmeter's high input impedance will also limit the voltage drop across the circuit under test and improve the measurement accuracy [32].

Figure 2-3 illustrates the measurement of a metal potential in an aqueous solution using a RE and voltmeter:

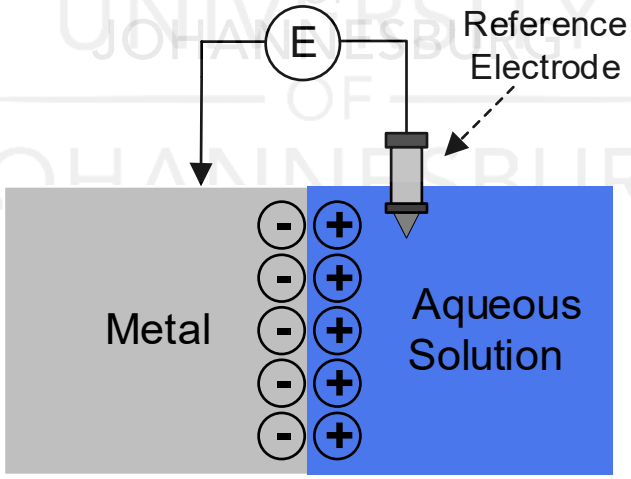

Figure 2-3 - Potential Measurement RE - Source: Adapted from [33]

The structure-to-electrolyte potential, or metal-to-electrolyte potential, is one of the critical measurements when determining the CP system's effectiveness. According to NACE, the structure-to-electrolyte potential is the potential difference between the

structure (metal) and RE. Each DC potential measurement needs to take the RE composition into account to ensure the DC potential readings are accurate [33]. CP pipe potential measurements in this study were measured with reference to a CSE and are denoted as "V_{CSE}".

2.2.4. Polarization

Callister refers to the process of polarization as the change in metal potential from its equilibrium state. Based on the Galvanic series, when two different metals are connected to form a corrosion cell, they will be at different potentials because the system is at a non-equilibrium state. The polarization process will force the anode potential to shift to the cathode potential and vice versa [28].

2.2.5. Corrosion Rate

According to Callister, the corrosion rate refers to the speed of a metal's decay, the corrosion penetration rate or the metal loss over time [22]. Various methods exist for expressing corrosion rates and for further reading, refer to Appendix A5.

2.2.6. Corrosive Environments

Callister suggests that corrosive environments include the atmosphere, aqueous solutions, soil, acids, bases, inorganic solvents, molten salts, liquid metals and even the human body [22]. Fontana et al. state that aqueous solutions can consist of fresh water, seawater or mine water. Furthermore, the process of industrialization in plants resulting in higher pressures, speeds and temperatures, also accelerates corrosion [27].

Callister further suggests that atmospheric corrosion is the most significant contributor to corrosion (based on a tonnage-per-year basis). Moist environments with dissolved oxygen are the primary corrosive agent, although sulphur compounds and sodium chloride can also contribute to the corrosion process [22].

Applicable to underground pipelines, soils can have different corrosive compositions due to non-uniform levels of oxygen, moisture, salt, acidity, alkalinity and bacteria. Due to the varying levels of soil composition, the corrosiveness of the soil will vary from one location to another [22].

2.2.7. Types of Corrosion

Corrosion can occur in both the inside or outside a metallic asset (depending on the asset type) [11]. According to Callister, eight distinct types of corrosion exists, namely [22]:

- Uniform Attack
- Galvanic Corrosion
- Crevice Corrosion
- Pitting Corrosion
- Intragranular Corrosion
- Selective Leaching
- Erosion-Corrosion
- Stress Corrosion

The ASM defines the following types of corrosion [35]:

- Atmospheric Corrosion
- Stray-current Corrosion
- Molten Salts Corrosion
- Liquid Metal Corrosion
- Microbiologically-induced corrosion [36]

This study focuses on external corrosion, such as stray-current corrosion. The NACE website can be consulted for a comprehensive description of each of the listed types of corrosion.

2.2.8. Corrosion Prevention

Because corrosion can occur on both the inside and outside of a metallic structure, proper corrosion prevention mechanisms are required, applicable to the location of corrosion. NACE suggests that investing in corrosion prevent technologies to reduce metal weight loss over time, which can reduce the risk of structure leaks and improve pipeline safety [11].

2.2.8.1. Inhibitors

Inhibitors applied in specific environments, and at low concentrations can reduce the corrosiveness of the environment. The inhibitor applied is specific to the metal type and composition of the corrosive environment [22].

2.2.8.2. Coatings

The primary corrosion prevention mechanism for metallic structures submerged in an electrolyte such as water or soil is a pipeline wrapping or coating. The coating aims to eliminate the contact between the anode, cathode and the electrolyte. Reduced contact between these elements can limit the corrosion of the cathode [37].

2.2.8.3. Cathodic Protection

Cathodic Protection techniques utilize the electrochemical properties of a metal structure to protect the metallic structure against corrosion by forcing the metal to become the cathode in an electrolytic cell when placed in an aqueous electrolyte [38]. Two main cathodic protection techniques exist, namely the sacrificial anode CP (SACP)system or the impressed current CP (ICCP) system [27].

2.2.8.3.1. Galvanic Anode Cathodic Protection

Galvanic coupling consists of two metals electrically connected and placed in a particular environment. One metal will lose electrons (called the anode) to protect the other metal (cathode). The metal sacrificing electrons, due to the oxidation reaction, is referred to as the sacrificial anode [22]. The galvanic series mentioned in a previous section indicates which metal will be the cathode and anode, respectively. The metal with the higher native potential will be the cathode, while the metal with the lower native potential, will be the anode [28].

Galvanic anodes supply a small current (<1A) in environments that do not require high magnitude currents. Furthermore, the use of a CP does not eliminate corrosion; it merely shifts it to the anode [18].

Galvanic anodes include installation at either a long section of a pipeline or at pipeline hotspots where a pipeline might not have a coating applied. The placement of the anode in proximity to the pipeline depends on various factors such as soil resistivity, current requirements, economic factors, pipeline coating and the anode metal type [18].

Figure 2-4 illustrates the concept of using a sacrificial magnesium anode for cathodically protecting an underground steel pipeline:

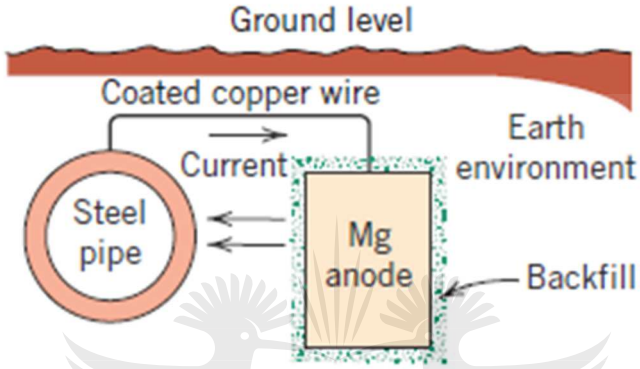

Figure 2-4 - CP using Sacrificial Anode - Source: Adapted from [22], [27]

2.2.8.3.2. Impressed Current Cathodic Protection

For underground pipelines, CP can prevent metal loss by reducing current flow in the electrolyte, by supplying a high-magnitude, reverse polarity, direct electric current, from an external source (such as a rectifier) [37]. A typical CP rectifier can produce an output voltage anything between 0V to a 100V and an output current from 0A to several hundred amperes. AC powered ICCP stations, require a DC rectifier to provide the CP current [18].

Figure 2-5 illustrates an ICCP system for an underground tank:

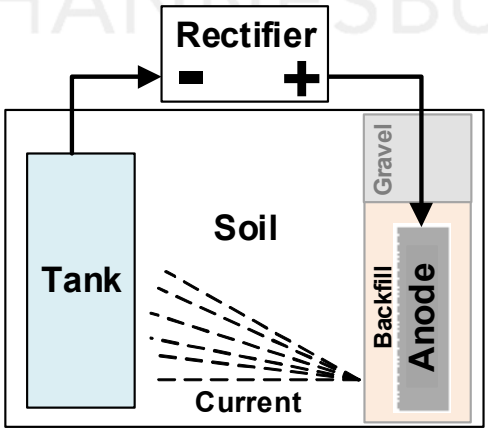

Figure 2-5 - ICCP System for Underground Tank - Source: Adapted from [22], [27]

The positive terminal from the rectifier is connected to the anode, while the negative terminal to the cathode (tank). Current flows from the anode to the cathode via the soil and completes the electrical circuit. This process shifts the cathode potential in a negative direction which reduces the corrosion rate of the metal [22]. The current applied to the tank aims to shift all anodic regions to cathodic regions in an attempt to reduce the cathode's corrosion rate [18]. In this example, the tank can also be a pipeline.

Peabody [18] illustrates a basic ICCP system below:

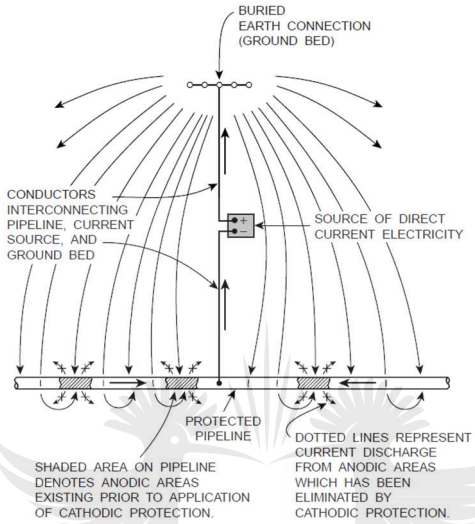

Figure 2-6 - Basic ICCP System - Source: Adapted from [18]

The cathodic areas on the pipeline initially collect current from the anodic areas on the pipeline. Upon installation of the anode ground bed and connection to the external power source, electrical current will start flowing from the anode ground bed, which will result in its consumption. A working CP system will eliminate current flow from the pipeline's anodic areas and hence stop the corrosion process [18].

This study focusses on ICCP systems, and section 2.3 discusses these systems in more detail. Appendix A6 compares the galvanic and ICCP systems typically used for CP system design.

2.2.8.3.3. Typical CP Equipment Used in Industry

From the literature available for CP implemented within the unique South African context (where high levels of stray current are present), the following CP equipment exists:

- Transformer Rectifier Unit (TRU) – The TRU is used for ICCP systems and powered by an AC source, solar power source or battery-powered [39], [40].
- Forced Drainage Unit (FDU) – The FDU is a powered ICCP unit which can also return current to a DC transit system from the pipeline to reduce stray current corrosion [39].
- Natural Drainage Unit (NDU) – The NDU is a passive unit which returns current to a DC transit system from the pipeline to reduce stray current corrosion. Also

used in cases where the pipeline is more electro-positive than the relevant interfering structure [39].
- AC Mitigation (ACM) – Reduce voltage spikes, AC density from AC induced corrosion, and ensure AC voltage is below the AC safety specification of 15VAC [39].
- Bonding – Connecting two different pipelines to ensure electrical continuity for CP current [41].
- DC-decouplers – Ground AC voltages or clamp DC voltages [39].
- Spark gaps – limit voltage surges [39].

2.2.9. Cost of Corrosion

The cost of corrosion is an economic decision that is based on the savings if a corrosion control system is implemented [27]. Fontana et al. suggested in 1987 that the annual cost of corrosion in the United States of America ranged between US8$ million and US126$ billion per annum. NACE estimated in 2013 that the global cost of corrosion was US2.5$ trillion per annum [4].

A trade-off between return-on-investment (ROI), maintenance and operation costs, and corrosion prevention costs to maximize the asset's life, while still making a profit. Formulas such as the net-present-value (NPV) and future value (FV), can be utilized the calculate the ROI over a fixed period [27].

According to NACE, the following factors affects the cost of corrosion [11]:

- Pipeline relocation due to excessive corrosion or construction activities close to the pipeline.
- Recoating the pipeline's external surface or application of corrosion inhibitors.
- Applying other corrosion prevention mechanisms such as anode backfill and isolation of electrical joints.
- Replacement of the pipeline before end-of-life.
- Plant shutdowns, product loss, reduced plant performance, product contamination, environmental impact and the cost of overdesign [21].

2.2.10. Risk Management Applicable to Pipeline Operations

Part of managing the cost of corrosion is the management of the corrosion risk. The risk level depends on the probability and consequence of a particular event [21]:

$$\boldsymbol{R = P \times C} \qquad \textit{2.3 – Risk Level Determination}$$

Where:

- R = Risk Level
- P = Probability of occurrence
- C = Consequence of occurrence

2.2.11. Section Summary

This section reviewed the basic principles of corrosion as well as corrosion mitigation techniques.

The next section delves into more advanced CP topics to build context for the scope of this study.

2.3. Cathodic Protection Monitoring And Management

This section investigates the applicable standards and regulations for the design and operation of CP systems to establish a framework for maintenance required by the relevant statutes and standards.

2.3.1. CP System Design, Operation and Maintenance

2.3.1.1. Statutory requirements According to 49 CFR PART 192

The Code of Federal Regulations (CFR), provides general and permanent rules as stipulated by the American Federal Government [42]. The safety requirements for the transportation of natural gas in pipelines are covered by standard 49 CFR PART 192.

This section discusses the applicable sections of the 49 CFR PART 192 statute to determine if absolute requirements exist.

2.3.1.1.1. External Corrosion Control

Permanently buried pipelines installed before 31 July 1971, must have an external corrosion protection system designed and installed, unless where sufficient tests indicate that the pipeline is not in a corrosive environment. The external corrosion protection system should include a protective coating and where applicable, a CP system. Buried pipelines installed before 1 August 1971 with an external coating, must have a CP system installed [43].

2.3.1.1.2. Cathodic Protection

The CP system must meet one or more of the followings requirements [43]:

I. *"A negative (cathodic) voltage of at least 0.85 volt, with reference to a saturated copper-copper sulfate half-cell. Determination of this voltage must be made with the protective current applied, and in accordance with sections II and IV."*
II. *"A negative (cathodic) voltage shift of at least 300 millivolts. Determination of this voltage shift must be made with the protective current applied and in accordance with sections II and IV. This criterion of voltage shift applies to structures, not in contact with metals of different anodic potentials."*
III. *"A minimum negative (cathodic) polarization voltage shift of 100 millivolts. This polarization voltage shift must be determined in accordance with sections III and IV."*
IV. *"A voltage at least as negative (cathodic) as that originally established at the beginning of the Tafel segment of the E-log-I curve. This voltage must be measured in accordance with section IV."*
V. *"A net protective current from the electrolyte into the structure surface as measured by an earth current technique applied at predetermined current discharge (anodic) points of the structure."*
VI. The CP system must be controlled and prevent excessive CP levels which can lead to cathodic disbondment of the pipeline coating.

2.3.1.1.3. Monitoring

Monitoring of the CP system must include the following [43]:

- Each pipeline's CP system must be tested at least once per calendar year and not exceeding intervals of 15 months
- If the above test is impractical, the tests can be split into portions of 10% per calendar year to cover the pipeline network over ten years

CP rectifiers and corrosion mitigation stations should be monitored as follows [43]:

- Every CP rectifier must be inspected every six months per calendar year and not exceeding 2½ months.
- Every reverse current switch, diode and interference bond must be inspected every six months per calendar year and not exceeding 2½ months. Other interference bonds must be inspected once per calendar year and not exceeding 15 months.
- Re-evaluation of unprotected pipelines should occur at least every three years, and CP applied to protect the pipelines in question.

2.3.1.1.4. Test Stations

Every pipeline must have sufficient test stations, or TP's, where electrical measurements are possible to determine the adequacy of the CP system [42]. According to Peabody, TP's should be at intervals less than 1.6km apart along the pipeline [18].

2.3.1.2. High-Level Conceptual Design of a CP System

A typical high-level design of a CP system includes a rectifier and downstream TP's:

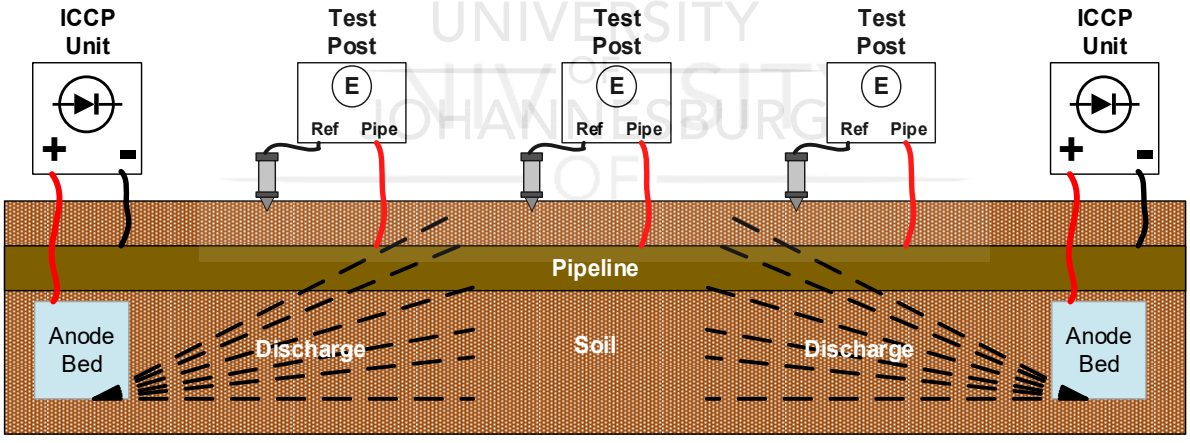

Figure 2-7 - High-level ICCP Design - Source: Adapted from [44]

The next section discusses the two factors applicable to this study when designing a CP system.

2.3.1.2.1. External Corrosion Control

Controlling external corrosion of a pipeline requires consideration of a CP system and pipeline coating during the pipeline design. Furthermore, the pipeline design should not lead to electrical shielding, which will effectively eliminate the CP current [11].

2.3.1.2.2. Corrosion Control Test Stations

NACE suggest the use of above-ground test stations to take voltage, current and resistance measurements. TP installations are typically at locations such as pipe casings; metallic crossings; insulation-joints (IJ); waterway, road and bridge crossings; valve stations; galvanic anode installations and ICCP installations [11].

2.3.2. NACE SP0169-2013 Standard

The NACE SP0169-2013 standard provides a guideline for the “Control of External Corrosion on Underground or Submerged Metallic Piping Systems”. This standard was initially published in 1969 and was reviewed by various task groups as advancements continued in the field. This standard focusses explicitly on factors relating to external corrosion such as electrical isolation, CP, stray current interference and insulating coatings [11].

Pipeline operators around the world follow this standard (and relevant sub-standards) to manage and control their pipeline anti-corrosion systems and reduce the cost of corrosion. Other regulatory standards include both the CFR and ASME standards (referenced throughout this document).

The following sections will cover the essential aspects of the NACE SP0169-2013 standard (and the relevant sub-standards) applicable to the scope of this study.

2.3.3. CP Monitoring

Determining the effectiveness of a CP system is achieved through measuring the pipe-to-soil potential, or CP pipe potential, at various intervals across the pipeline and comparing the measurements with the NACE standard. The NACE SP0169-2013 standard provides three criteria's to determine if the CP potentials meet the set standard [45].

To measure the voltage across the pipe and soil interface, a RE is required. The standard RE used for underground pipelines buried in the soil is a CSE, which can be permanent or stationary [18].

Some of the most critical elements that affect the accuracy of measured CP pipe potential are the measurement setup, IR drop and temperature effects. Discussed in the sections below are the factors affecting the pipe-to-soil potential and the NACE criteria for determining the effectiveness of the CP system.

2.3.3.1. Measurement Techniques

2.3.3.1.1. Instrument and Measurement Guidelines

The NACE TM0497-2018 standard provides instrument and measurement guidelines to ensure the recording of accurate electrical potential readings. Considerations include the meter selection (analogue or digital); channel input impedance; sensitivity, analogue-to-digital converter (ADC) speed, instrument accuracy, readout resolution, sampling rate; AC and RF rejection; and environmental limitations such as temperature and humidity [46].

2.3.3.1.2. Pipe-to-Soil Measurement Guidelines

NACE suggest a minimum input impedance of 10MΩ per input channel to reduce voltage drops and measurement errors. Verification of the measurement accuracy can be done with two independent meters or the same meter with two different input impedances and comparing the measured potentials. NACE further suggest that meters be calibrated annually and taken out of service if the measurement is not accurate [46].

2.3.3.1.3. Pipe-to-Soil Measurement Techniques

With CP applied to a pipeline (and no dynamic stray current exists), one can expect a negative pipe-to-soil potential (V_{CSE}). The meter test lead polarity will affect the sign displayed in front of the measured potential. NACE recommends connecting the COM terminal to the CSE and the VOLT terminal to the pipeline to ensure that the polarity of the measured value is correct [46].

Figure 2-8 illustrates the recommended multimeter terminal connections [46]:

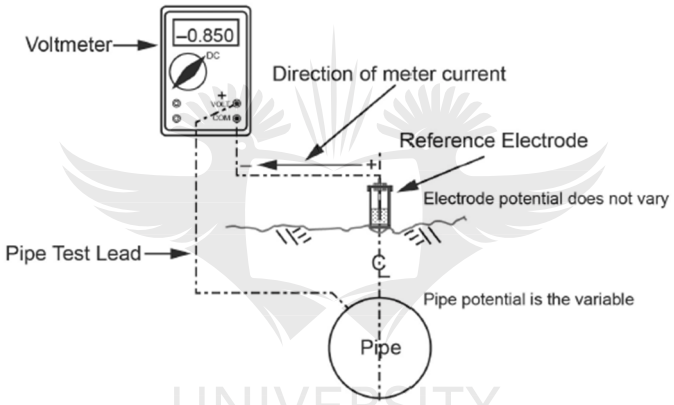

Figure 2-8 - Instrument Connection (Recommended) – Source: Adapted from [46]

2.3.3.2. Factors Affecting the Measurement Accuracy

2.3.3.2.1. General

General factors affecting the measurement accuracy includes the multimeter, and the connection leads and whether they are in a proper working condition. Furthermore, loose or corroded connection points can also affect the measurement circuit [46].

RE's can also impact the measurement accuracy due to contamination, high contact resistance, exposure to sunlight, blockage of the porous plug or improper placement with relation to the pipeline [46].

Electromagnetic Interference (EMI) from overhead AC powerlines, as well as uncontrolled DC stray currents and telluric currents, can also affect the measurement accuracy. Furthermore, NACE also suggests that long periods of depolarization will also affect measurement accuracy [46].

2.3.3.2.2. IR Drop

According to Ohm's Law, the IR drop is the voltage across a resistance. In a typical pipe-to-soil measurement, various resistances are present in the circuit, which contributes to the total IR drop. When taking potential measurements, the voltage drop across the pipe-to-soil interface needs to be taken into account to reduce the error in the reading taken. The IR drop causes the potential to shift more negative, which might falsely indicate that the CP applied is sufficient [44].

Several measures can be applied to reduce the IR drop, such as reducing cable lengths, placing the electrode close to the pipeline, use of a coupon or taking instant-off potential measurements with a current interrupter at a rectifier. Dimond and Ansuini further suggest that the input impedance should be at least 10MΩ at all times during a measurement to prevent RE polarization [34].

Holtsbaum suggests two methods to eliminate or reduce the IR drop from a potential measurement [45]:

- Place the reference electrode as close as possible to the pipeline
- Interrupt the CP current and take a measurement, this will result in 0A and thus a 0V IR drop based on Ohm's Law

In an experiment conducted by Holtsbaum, measurement errors result due to neglecting the influence of the IR drop. In this specific example, the measurement error is 200mV, which will be false reporting of the actual value of the potential measurement [44]. Holtsbaum further indicates that the IR drop will also increase due to misplacement of the RE about the pipeline (due to the current and structure resistance) [44].

Where the IR drop is not known, industry experts suggest using correction factors, perform distance extrapolation, using current interrupters or coupons [47].

2.3.3.2.3. Temperature Effects

Temperature affects the RE and requires compensation for the potential measurement above and below 25ºC [45].

The temperature compensation formula is [45]:

$$\boldsymbol{E_t = E_{25°C} + k_t \times (\mathrm{T} - 25°\mathrm{C})} \quad \textit{2.4 – Temperature Compensation}$$

Where:

- E_t = Reference potential at temperature *t*
- $E_{25°C}$ = Reference potential at 25ºC
- k_t = Temperature coefficient
- T = Temperature in ºC

Common practice only requires temperature correction if ten Degrees-Celsius above and below the 25ºC guideline [11]. NACE provides temperature coefficients for the different RE's [11]:

Common Reference Electrodes and Their Potentials and Temperature Coefficients					
Reference Electrode	**Electrolyte Solution**	**Potential at 25°C [77°F] (V/SHE)**	**Potential at 25 °C [77°F] (V/CSE]**	**Temperature Coefficient mV/°C (mV/°F)**	**Typical Usage**
$Cu/CuSO_4$ (CSE)	Sat. $CuSO_4$	+0.316 [76]	0	0.9 (0.5)[76]	soils, fresh water
Ag/AgCl[(A)] (SSC)	0.6 M NaCl (3 ½%)	+0.256[77]	–0.06	– 0.33 (0.18)[77]	seawater, brackish[(B)]
Ag/AgCl[(C)] (SSC)	Sat. KCl	+0.222[78]	–0.094	– 0.70 (0.39)[78]	---
Ag/AgCl[(C)] (SSC)	0.1 N KCl	+0.288[79]	–0.028	– 0.43 (0.24)[79]	---
Sat. Calomel (SCE)	Sat. KCl	0.2441	–0.072	– 0.70 (0.39) [10]	water, laboratory
Zn (ZRE)	Saline Solution	–0.79 ± 0.1[65]	–1.1 ± 0.1[65]	---	seawater
Zn (ZRE)	Soil	–0.80 ± 0.1[65]	–1.1 ± 0.1[65]	---	underground

(A) Solid junction.

(B) Potential becomes more electropositive with increasing resistivity. See nomograph for correction in waters of varying resistivity in NACE SP0176,[10] or see reference 77.

(C) Liquid junction.

Table 2-1 NACE RE Temperature Coefficients - Source: Adapted from [11]

2.3.3.2.4. Stray Current Interference

Wang et al. define stray currents as any current that does not follow its intended path [48]. Stray current corrosion, which consists of both AC and DC, is a significant cause of corrosion because it forces the structure to become anodic. DC interference poses the most significant risk due to the current pickup on the pipeline [46]. This interference can directly translate to metal loss, and estimations suggest that a continuous DC discharge of 1A can result in a metal loss of 10kg over one year [49]. Telluric currents can also cause stray currents due to geomagnetic fluctuations [46].

Structure-structure corrosion caused by current flow in the structure causes a potential gradient, while earth-current corrosion results from current flow in the electrolyte [49].

Typical causes of DC interference includes foreign pipelines or structures [49]. Another source of stray current apparent in the South African context, is current from DC transit systems, where the electrical current leaves the rail and jumps onto the pipeline. If no current return path exists to the rail, corrosion can increase at a rapid rate [41].

Figure 2-9 illustrates stray current interference from a DC transit system:

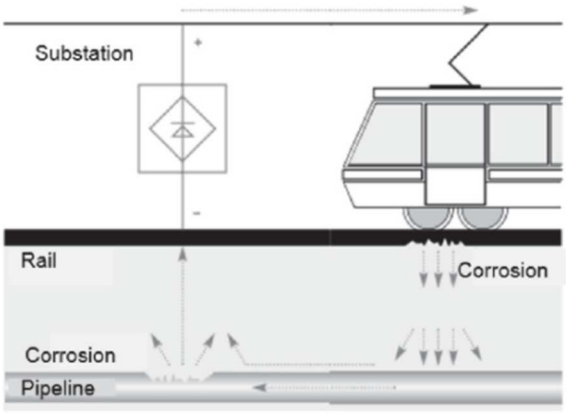

Figure 2-9 - DC Transit System Stray Current - Source: Adapted from [50]

Calculation of the current density can be performed with formulas presented in Appendix A7.

2.3.3.3. NACE SP0169-2013 CP Criteria's for Steel Pipelines

The NACE SP0169-2013 suggests three criteria to determine the adequateness of CP applied to a steel pipeline, namely, instant-on potential criteria, instant-off potential criteria and 100mV cathodic polarization criteria [11]. All three criteria's are discussed below, as well as the use of coupons.

2.3.3.3.1. Instant-On Potential Criteria

The instant-on potential criteria refer to a pipe-to-soil potential less than -850mV$_{CSE}$, when measured with reference to a CSE, and when CP current is applied. The measurement includes the polarized potential and the measurement circuit's voltage drop. Since the voltage drop is unknown, an analysis of the CP system's performance over time is also required to meet this criterion [11].

2.3.3.3.2. Instant-Off Potential Criteria

The instant-off potential criteria refer to a pipe-to-soil potential less than -850mV$_{CSE}$ when measured with reference to a CSE without CP current applied. NACE suggests the use of a current interrupter to momentary switch off the ICCP unit, to take an instant-off potential measurement within three seconds. This measurement will only reflect the polarized potential and will exclude any voltage drops (except if adjacent ICCP units still supply current). The interruption of CP current will result in transient depolarization of the pipeline, which removes the IR drop. Waiting too long to take the measurement can result in erroneous potentials. Synchronization of current interrupters is a complex process because all ICCP units should switch off at the same time [11].

2.3.3.3.3. 100mV Cathodic Polarization Criteria

The 100mV Cathodic Polarization criteria refer to a voltage difference of at least 100mV between the pipeline's native potential and corrosion potential (E_{CORR}) [11].

Figure 2-10 illustrates the potentials for the three different criteria's:

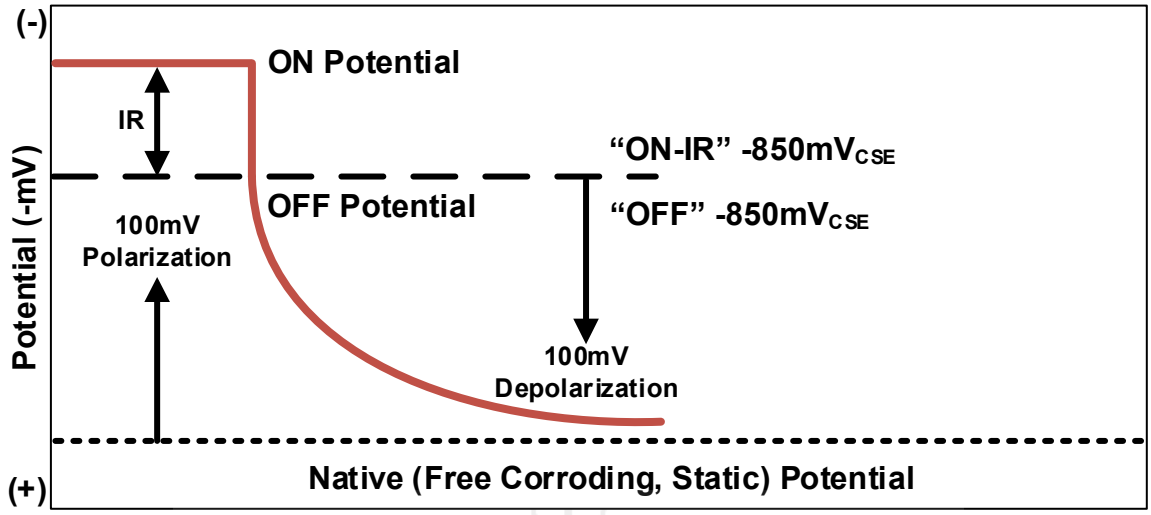

Figure 2-10 - Typical Depolarization Curve - Source: Adapted from [28]

2.3.3.3.4. Coupons

The NACE/ANSI standard, RP0104-2004, provides guidelines for using coupons when determining the effectiveness of CP systems through an accurate polarized potential measurement. Coupons are devices installed next to the pipeline that simulates a coating holiday (bare area) on the pipeline [51].

The coupon-to-electrolyte potential is measured by disconnecting the coupon and taking the instant-off potential. The coupon can also remain disconnected over time to measure the polarized potential. Satisfying the NACE 100mV Cathodic Polarization criteria requires the use of both the coupon-to-electrolyte and polarized potential. Important to note is that the pipe-to-soil potential and coupon-to-electrolyte potential will differ at the same location, due to the presence of an IR drop in the pipe-to-soil potential [51].

Coupons also allow for the measurement of coupon current magnitude and direction. Current discharge from the coupon indicates inadequate CP protection and a possible risk of corrosion. Current pickup can mean adequate CP protection is in place if no other current interference is present. The current density can also be estimated based on the surface area of the coupon. The main advantage the coupon offers is an IR-free potential, but the once-off installation costs are high [51].

2.3.3.4. Other Corrosion Measuring Techniques

Ameh et al. [57], suggest four other corrosion survey methods:

1. Potential Surveys – Includes the two primary techniques, direct current voltage gradient (DCVG) to detect coating holidays and close interval potential surveys (CIPS) to measure the pipe-to-soil potential along the pipeline [52].
2. Corrosion Coupons – as discussed in section 2.3.3.3.4
3. Bacteria – Includes the monitoring of bacteria present which accelerates microbial corrosion [52]

4. Intelligent Pigging – In-line monitoring method to detect the metal wall loss, lamination, cracks and ovality [52].

Appendix A8 lists the corrosion measuring techniques suggested by Holtsbaum.

2.3.3.5. ICCP Rectifier Maintenance

Holtsbaum provides maintenance and operation guidelines for ICCP rectifiers and is discussed in the sections below [44].

2.3.3.5.1. Rectifier Components

An ICCP rectifier typically consists of the following elements [44]:

- AC supply with primary transformer, surge protection and circuit breaker/s
- Transformer with related tap options and fuse protection
- DC rectifier elements (typically a full-wave bridge rectifier), fuse and surge protection
- Instrumentation to monitor AC voltage and DC voltage and current

Figure 2-11 provides a basic overview of an ICCP rectifier [44]:

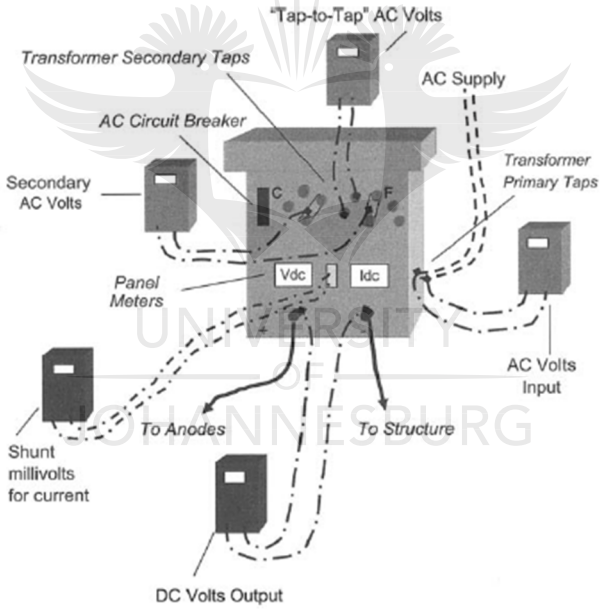

Figure 2-11 - Basic ICCP Rectifier - Source: Adapted from [44]

Appendix A9 illustrates a flowchart for ICCP rectifier fault finding.

2.3.3.5.2. Adjusting Rectifier Settings

Holtsbaum suggests that rectifier setpoints only be changed if the cause for the adjustment is known to prevent other anomalies from occurring. Rectifiers generally operate in the following modes and should be adjusted according to the active operational setting [44]:

- Constant Current Mode – The rectifier maintains a constant current output to compensate for resistance changes in the anode. The voltage will vary based on resistance changes to keep the current constant [44].
- Constant Potential Mode – The rectifier will maintain a constant potential between the pipe and reference electrode by adjusting the current of the rectifier. Suppose the pipe-to-soil potential goes more electro-positive, the current increase, and vice versa [44].
- Constant Voltage Mode – The rectifier maintains a pre-set output voltage [40].

2.3.3.5.3. Inspections

Holtsbaum suggests monthly routine inspections to reduce the time whereby a rectifier will not provide CP to a pipeline. These inspections should prioritize capturing the CP pipe potential, rectifier voltage rectifier and current. If remote monitoring exists, the data integrity is high, and no system errors exist, one can opt to increase inspection intervals. Annual inspections should include the calibration of panel meters at each rectifier, inspection for hot terminals and determining seasonal anode bed changes to prevent the rapid deterioration of anode beds (resistance will increase) [44].

Where an ICCP unit is offline, reactive maintenance aims to reduce the impact of insufficient supply of CP to the pipeline [44].

The use of both condition-based and predictive maintenance will aid in the maintenance response required and reduce the cost to maintain the CP system.

2.3.3.6. Remote Monitoring

Process Control Systems (PCS) enables the monitoring and control of remote sensors from a centralized location. As a sub-system of a PCS, SCADA systems, monitors and control remote sensors over a large geographic area through data acquisition, networked data communication, data presentation and control. Typical SCADA deployments are in manufacturing industries, pipeline operators, municipalities (bulk/ wastewater management) and electrical utilities. Telemetry devices enable data collection over long distances through communication technologies such as General Packet Radio Services (GPRS) or satellite communication [12].

Typical telemetry-based SCADA deployments consist of multiple sensors read by a Remote Terminal Unit (RTU) at a remote location. The RTU reports back data to a Master Terminal Unit (MTU) which initiates all communication to and from the RTU stations. The MTU-RTU communication backbone typically consists of fibre optic networks, licensed microwave networks or GPRS modems. The SCADA system receives data from the MTU using a protocol driver and presents the received data using a graphical user interface (GUI). Based on the design, the SCADA operator can

control some of the equipment interfaced into the RTU station [12]. Telemetry systems enable CP system monitoring of large pipeline networks.

Yang suggests that frequent monitoring of a CP system can enhance the build-up of a database of CP data for optimization of the CP system operation [26]. Remote monitoring of long pipelines aims to collect data from a variety of sensors to enable continuous monitoring of the pipeline. Effective remote monitoring can eliminate the requirement for expensive field surveys every few years and should aim to monitor corrosion and system failures in real-time [53].

Peratta et al. proposed a CP remote monitoring architecture by using RTU's with Global System for Mobile Communications (GSM)-enabled communication to monitor ICCP units on a transmission pipeline in Europe. The RTU's transmit data to a central server for data collection, processing and sending of alert messages. The main aim of the remote monitoring system is to collect CP pipe potentials and provide this data to a secondary software application (BEASY) for pipeline condition evaluation [53].

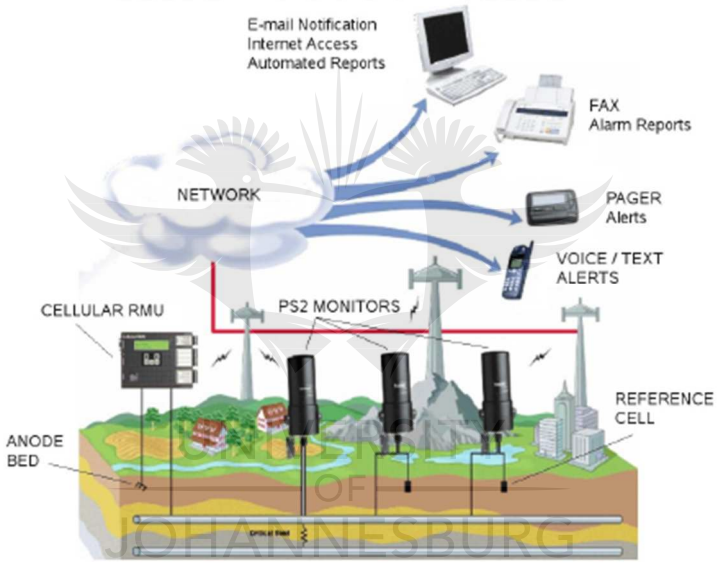

Figure 2-12 - CP Remote Monitoring Solution - Source: Adapted from [53]

Hoppe et al. further suggested that the use of a CP remote monitoring unit (RMU) should enable the detection of a rectifier failure, allow for rapid identification of CP anomalies; flag stray current areas; and provide an early warning for polarization cell problems. Visualization of data on a dedicated GUI can enable control room operators to react to alarms as it occurs. In this system architecture, the use of RTU's with an internal storage memory of about 20 days prevents data loss should the communication to the RTU fail [54].

From Hoppe et al.'s case study, he was able to determine the following from the RMU system [54]:

- Rectifier voltage or current loss
- Overvoltage of a polarization cell
- Erratic shifts in CP pipe potential

- Significant drops in CP pipe potential

Although the deployment of an RMU system is an expensive CAPEX cost, the cost of reducing the scheduled maintenance requirements offsets this initial CAPEX investment [54].

In both of the above systems, relational databases stores received data, that allows for further processing (knowledge extraction). SCADA systems enable the visualization of received data and forwards collected data to a database [12]. The use of data loggers are also becoming more popular in the CP industry, and data from TP's along the pipeline can be recorded continuously and sent to a central server using GSM communication.

In more recent research, Abate et al. suggest the use of a 169MHz M-Bus networked CP RMU system, that utilizes a fuzzy logic controller to send impressed current values back to a CP receiver (typically at the ICCP unit) [55].

The figure below illustrates the M-Bus network, with MP as the unit with the power supply and MP1 and MP2 being remote nodes:

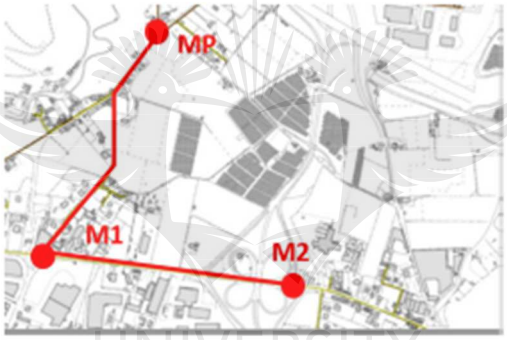

Figure 2-13 - M-Bus CP Monitoring System - Source: Adapted from [55]

Kara et al. deployed a CP monitoring system based on linear wireless sensor networks (LWSN) that allows for long-range, low bit-rate, bi-directional communication between sensor nodes for long pipelines [56].

ICCP rectifier monitoring typically includes the following:

1. AC Power Supply
2. Circuit breaker positions
3. Surge protection and fuse monitoring
4. The rectifier output voltage, current and frequency
5. Pipeline AC and DC potential
6. Diagnostic measurements for the monitoring system

2.3.4. Pipeline Integrity Management System

NACE suggests that pipeline operators invest in a PIMS to effectively manage pipeline hazards that includes pipeline technical hazards, processes and procedures, risk assessments and management programs or systems [57]. Data collected from CP systems usually drive the decision-making process.

NACE suggests that the current level of protection of a pipeline consists of data from various sources such as pipe-to-soil potentials, rectifier surveys, daily operating records (from a SCADA system), line inspections, inline-inspections and ECDA tests [57].
The ASME B31.8 standard suggests data collection from various sources (to the same pipe location reference) can be consolidated in a single system to improve the effectiveness of the integrity management system [58].

The CSA Z662-2007 standard [59] stipulates that pipeline operators should implement measures for safe pipeline operations and have effective product loss procedures in place. Pipeline safety includes management buy-in, optimised organisational structures, resource management procedures and evaluation and training and education programs [57].

The PAS 55-1:2008 standard provides general guidelines to manage assets and can enable an organization to effectively manage their assets in terms of performance, risks and costs that spans the asset's lifecycle. Performance monitoring of the asset can provide leading and lagging indicators. The former being a set of metrics based on historical events (such as incidents) and the latter being an indication of performance [60].

Although PIMS does not form part of the study's scope, managing pipeline corrosion and operations require an integrative approach.

2.3.5. Corrosion Management System

NACE performed an impact study on the implementation of a CMS framework for a pipeline which is concerned with cost-effective pipeline operations. The proposed CMS system is concerned with various aspect of the organisation, such as management, asset management, quality management, safety management and environmental management. The proposed CMS aims to follow multiple international standards to ensure the sustainability of pipeline operations.

Figure 2-14 presents the proposed CMS building blocks [61]:

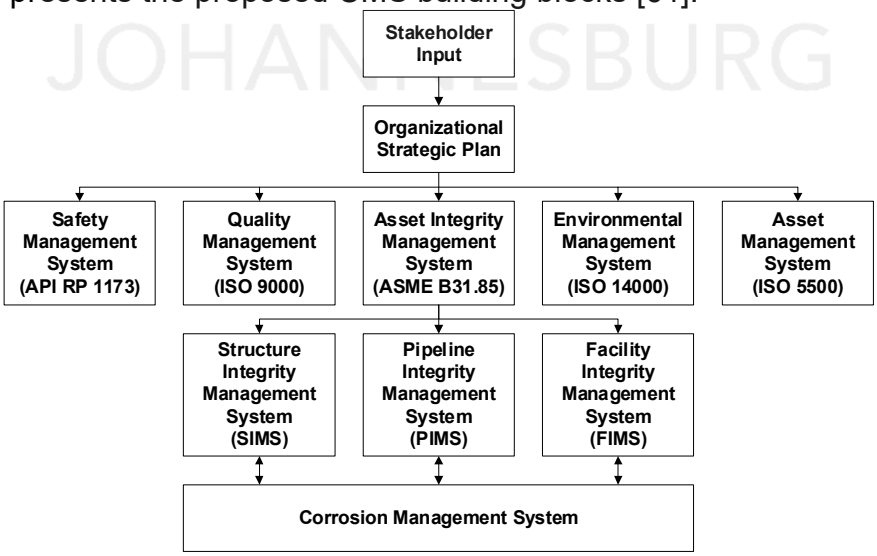

Figure 2-14 - CMS Building Blocks - Source: Adapted from [61]

The CMS is not part of the study's scope but also suggests that an integrative approach, similar to PIMS, is necessary for pipeline corrosion management. The proposed predictive maintenance approach of this study can potentially integrate with an existing CMS.

2.3.6. Section Summary

This section reviewed the literature for designing, operating and maintaining CP systems. The NACE SP0169-2013 criteria's provides the foundation of the statistical analysis for this study. Remote monitoring of CP systems provided a background as to how CP data retrieval works for remote sites. This section concluded with a glance of PIMS and CMS frameworks.

The next section delves into reliability engineering principles and maintenance strategies.

2.4. Reliability Engineering Principles

The International Organisation of Standards (ISO) defines an equipment failure as the "termination of the ability of an item to perform a required function" and prognostics as the "analysis of the symptoms of faults to predict future condition and remaining useful life"[62]. The latter is of importance for the scope of this study and the equipment failure mode, failure rate and fault progression sequence drives the equipment state prediction.

O'Connor et al. suggest that reliability is the probability that an asset will perform at optimal condition for its intended use. Reliability definitions vary per industry, but the general description is the number of equipment failures over time. Reliability refers to the time to system failure, whereas availability refers to the total system uptime and maintainability is the ability of a machine to be restored to a state where it can perform for its intended use [62], [63].

The three equipment failure rates associated with reliability engineering is the burn-in -, useful life - and wear out rates as per the famous Weibull bathtub curve. The curve indicates that the failure rate is initially high and decreases over time (burn-in phase) where after the failure rate stays constant over a period of time (useful life) phase. At the end of the equipment life, the failure rate increases again (wear-out phase) until the equipment fails completely. Failure patterns are required to estimate the remaining useful life of a system or equipment [63].

The Weibull-curve below illustrates the equipment operating life:

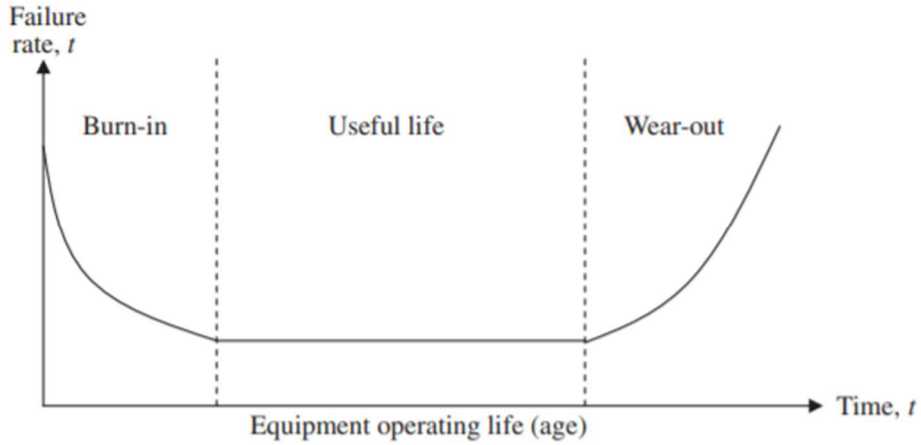

Figure 2-15 - Weibull - Bathtub Curve - Source: Adapted from [64]

Typically statistical measures include the equipment failure rate, mean time between failures (MTBF), mean time to failure (MTTF) and the mean time to restore (MTTR). Statistical analysis requires consideration of both equipment failure rates and the relevant maintenance approach [63]. The root-cause analysis (RCA) can provide insight into the cause of system failure (failure mode) and possible mitigation actions [62].

Standard techniques for modelling the reliability of a system exists such as physics of failure (POF), fault-tree analysis (FTA), failure modes and effects analysis (FMEA), and reliability block diagrams [63].

Although there is abundant literature available on reliability engineering topics, limited to no research is available for the reliability of CP systems which indicates minimal research performed on CP systems in the past. This study aims to use some of the existing reliability principles and apply them to CP systems.

The sections following discuss the literature available for condition monitoring systems, different maintenance strategies and statistical analysis of reliability data.

2.4.1. Condition Monitoring Systems

Condition monitoring (CM) determines the health of a system in operation and usually consist of a set of sensors that monitors various aspects of the system. Faults that occur in a system can either be a hard or abrupt fault or a soft fault which occurs over some time. Soft faults can model future equipment states, whereas hard faults result in definite equipment failure and are less probable to be predicted [65].

CM systems focus in diagnostics and prognostics, the former describing the fault state of the machine (either from previous operating data, alarms, or comparison) and the latter the progression of existing and future faults. Prognostics indicate the estimated time to failure (ETTF). CM systems alarms can trigger messages to alert personnel of anomalies [66].

The ISO 17359:2018 standard suggests an audit focussing on the reliability and criticality of the system in question, which can also aid at improving the performance of existing CM systems [66].

ISO recommends performing a baseline cost-benefit analysis to establish accurate KPI's to measure the effectiveness of a CM programme. Costs to consider include the life cycle costs, the cost associated with production losses or rework, consequential damage and warranty and insurance costs [66].

The standard system architecture of CM systems based on the Open System Architecture for Condition-Based Maintenance (OSA-CBM) framework, consists of the following [65]:

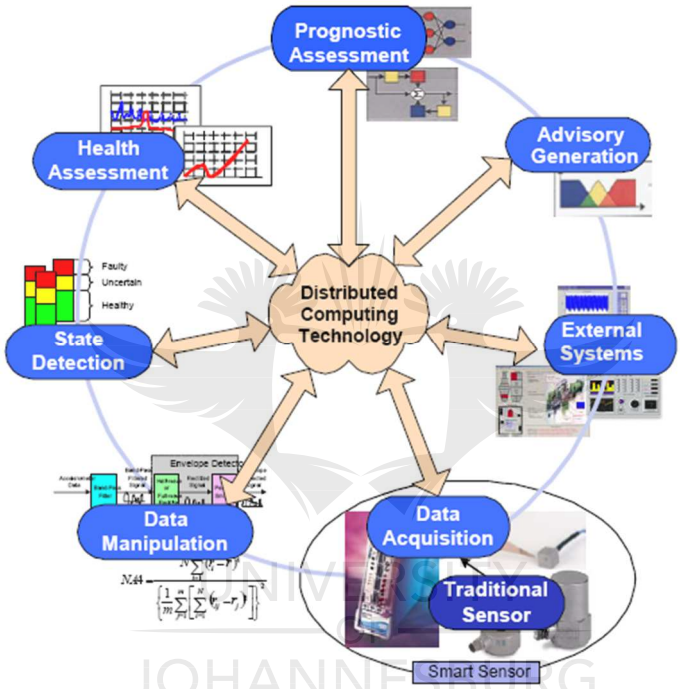

Figure 2-16 - CM System Architecture (OSA-CBM) - Source: Adapted from [65]

The data acquisition module is responsible for collecting data from sensors (either traditional or smart sensors), while the data manipulation module transforms the collected data using CBM feature extraction algorithms. The state detection module compares the extracted features to baseline operating conditions in the state detection module and issues an alert to plant personnel. The system health assessment is continuously performed to determine the diagnostic state of a system and considers both current and previous health assessments and maintenance records. The prognostic assessment module projection of the future health state and typically includes the remaining useful life (RUL) calculation. The advisory generation module provides recommended actions to achieve mission objectives and decision support, while external systems integration provides historical maintenance data or other data required by the CM system [67].

Fault prognostics consist of three standard approaches, namely a data-driven approach, a model-based approach or a hybrid approach of the model and data-driven approaches. The data-driven approach collects data from sensors and extracts features for RUL prediction. This approach is more simplistic and cost-effective to implement in comparison to the model-based approach, but a trade-off exists in terms of prediction accuracy. Model-based approaches are more accurate and are developed based on the physical (mathematical) characteristics of the system, but is costly to implement and requires the development of a degradation model [68].

Classification of CM data fits into two categories and is either condition monitoring data (equipment operating or health state data) or event data (equipment failure and maintenance). Event data indicating failure mechanisms, also referred to as lifetime data, can be used to determine equipment survival functions. The popular Kaplan-Meier (KM) technique can predict the RUL [69].

The proportional hazard model (PHM), linear regression (LR) or support vector machines (SVM) are popular models for calculating the RUL in statistical-driven approaches [69].

2.4.2. Maintenance Strategies

Maintenance strategies are employed to ensure any system remains operational. These strategies include actions such as identification, equipment repairs and replacement and inspections. Maintenance strategies require executable instructions and tactical plans [70].

Gackowiec [70] created a maintenance classification matrix based on the available academic literature in various academic databases:

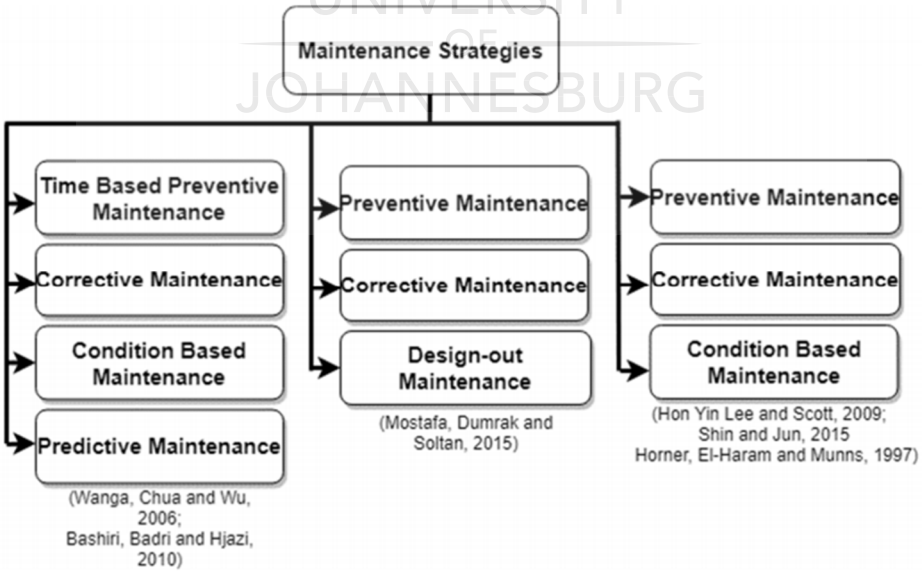

Figure 2-17 – Maintenance Strategies - Source: Adapted from [70]

Although the NACE standard provides standard maintenance and inspection requirements of CP equipment, reviewing some appropriate maintenance strategies

in the field of reliability engineering can serve as use-cases for maintenance of CP systems:

2.4.2.1. Preventative Maintenance

Preventative maintenance (PM) aims to keep a system operational. Maintenance activities include planned inspections, cleaning, calibration or testing of systems at regular intervals [63]. Some literature also refers to PM as time-based maintenance [70]. The NACE SP0169-2013 standard sets the minimum criteria for preventative maintenance of different CP equipment discussed in section 2.3 [11].

2.4.2.2. Corrective Maintenance

Corrective maintenance, or run-to-failure maintenance, aims to restore system operation after equipment failure has occurred [63]. Corrective maintenance strategies lead to high downtime of equipment and uncontrolled maintenance costs [64].

2.4.2.3. Condition-based Maintenance

Condition-based maintenance (CBM) monitors a set of data from equipment to detect abnormalities and aims to prevent equipment failure. Some literature refers to CBM as predictive maintenance (PdM). CBM consists of either a periodic inspection and replacement (PIR) strategy or a quantile based inspection and replacement strategy (QIR) [70]. The main aim of this strategy is to reduce costs and ultimately, machine downtime [64]. Based on the remote monitoring of CP systems, Hoppe et al.'s case study illustrates the ability to detect failures defined explicitly for a CP system [54].

Various industries utilize CBM systems to predict equipment failures in applications such as vibration monitoring, oil analysis and lubricant monitoring, and sound or acoustic monitoring. Although not used frequently, condition monitoring of electrical circuits can detect isolation issues, circuit shorts and broken motor rotor bars. CBM focusses on two areas, namely diagnosis and prognosis. The former being concerned with provided early warning signs of equipment failure, while the latter aims to predict when the failure will occur [64].

Various ISO standards are available, providing the guidelines for the design and implementation of a condition-based monitoring system and determining machine diagnostics [71]. Decision making consists of current condition evaluation-based (CCEB) and future condition prediction-based (FCPB). CCEB is concerned with the current condition of the equipment and will determine the current maintenance required while FCPB predicts the future trend of equipment failures [64].

Data analytics are usually used to determine the equipment predictions and can consist of multiple techniques such as neural networks, genetic algorithms, feature extraction [64]; Markov models, Bayesian models, Monte-Carlo Simulation and network models [72].

2.4.2.4. Time-based Maintenance

Time-based maintenance (TBM) strategies consider the failure rate of equipment to determine the maintenance schedule. Statistical modelling enables prediction of the

equipment failure rates, identification of equipment failure patterns/trends and the calculation of the MTTF based on the bathtub curve [64].

2.4.2.5. Risk-Based Maintenance

Risk-based maintenance (RBM) aims to decrease the probability and consequence of equipment failures by optimising maintenance planning and execution [73]. Risk-based inspections (RBI) was proposed for subsea pipelines by Singh and Markeset to reduce the cost of consequence and maintenance. For this approach, a fuzzy methodology estimates corrosion rates [74]. Xu presented a predictive maintenance strategy that incorporates both risk and equipment conditions by using a probabilistic inference with bucket elimination design [73].

2.4.2.6. Reliability-Centred Maintenance

Reliability-centred maintenance (RCM) focusses on the system function and not on the system hardware. RCM aims to reduce maintenance costs by prioritizing maintenance activities that will affect the system function[75].

2.4.3. Reliability Evaluation

ISO suggests that a confidence interval is required when evaluating reliability data, to ensure the calculated reliability is accurate. The degree of data reliability is dependent on facts and the sample size [66].

Reliability data analysis requires the use of statistical measures to enable decision making. The probability density function (pdf) of a dataset determines the distribution of the dataset and can consist of unimodal or multimodal distributions. O'Connor et al. suggest four factors to be determined when describing a pdf namely, central tendency (grouping of data), the variation of the dataset, dataset skewness (symmetry or lack thereof), and the kurtosis of the dataset (peaks present) [67]. Evaluation of the reliability of equipment can consist of a combination of probability theory and statistical analysis [63].

For an evaluation of systems that uses data, the reliability and integrity of the data should be high. Various measures exist to test the consistency of different data sets and includes Cohen's kappa for nominal data, Pearson's correlation coefficient, and the percentage of agreement and index of concordance [76].

2.4.4. Section Summary

This section evaluated reliability engineering principles that can potentially inform the research design of this study.

The next section evaluates the evolution of data analytics and typical data modelling approaches.

2.5. Data Analytics

Ramasubramanian and Singh refer to the importance of statistics in data analysis and points to the following statement: "Statistics as the science of learning from data, and of measuring, controlling, and communicating uncertainty is the most mature of the data sciences" [77].

ML is a data analytics method that teaches a machine how to process data more efficiently, especially in a scenario where a human cannot identify a pattern from the dataset [78]. Further advancement of technology led to artificial intelligence (AI), which is a core feature of robotics. AI refers to “the science and engineering of making intelligent machines” [77]. Some literature also refers to ML as a subset of AI.

Data mining, also referred to as Knowledge Discovery in Databases (KDD), seeks to retrieve information and knowledge from large data sets which can be incomplete, random or noisy. Data mining differs from typical data analysis because it seeks to mine information and discover knowledge [77].

This section describes some of the data analytics methods available for the evaluation of data sets and model predictive algorithms.

2.5.1. Probabilistic Methods

The probability of an event describes the likelihood of reoccurrence under a series of circumstances. Probability is a value between 0 and 1, the former indicating the event is never expected to occur, and the latter having a high likelihood of occurring [79].

Three general distributions include the binomial distribution (consist of two mutually exclusive events), Poisson distribution (event occurs within a timeframe) and distributions for continuous variables (such as the Gaussian or normal distribution). The mean (centre value of the distribution) and the standard deviation (difference in values from the mean) describes the characteristics of a continuous distribution [79].

Probabilistic approaches enable decision making, where uncertainty exists. Decision trees, influence diagrams and reduction algorithms are popular methods for modelling the system. Influence diagrams provide a graphical representation for decision making, where a high level of uncertainty exists to determine the conditional independence and its required data [80].

Clustering techniques groups the same variable in various other groups for efficient evaluation and computation of joint distributions (applicable to probabilistic inference problems) [80].

2.5.1.1. Predictive Modelling Overview

Kuhn and Johnson define predictive modelling as the process to define a mathematical model to make accurate predictions. The methodology of model building consists of data splitting; identifying predictors; estimating performance using quantitative statistics; evaluating various models, and selecting the most appropriate model. Data transformation is required in some scenarios to handle issues with predictors such as resolving skewness; centring and scaling data; removing outliers in the dataset; handling missing data and making changes to predictors (adding, removing or binning) [81].

ML techniques present various issues in the format of over-fitting (ML model learns the noise of the model as well, and the accuracy of the prediction is affected). Model tuning can prevent over-fitting. Resampling techniques can also be employed to fit the model and determine the efficacy of the model by splitting data sets [81].

Figure 2-18 represents a typical ML model tuning process flow:

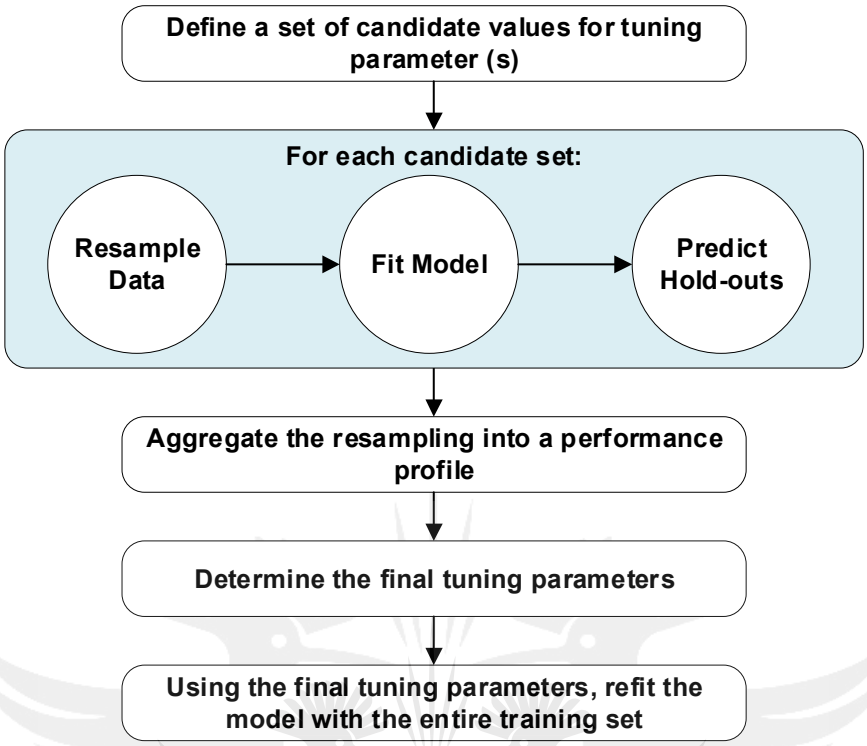

Figure 2-18 - Model Tuning Process - Source: Adapted from [81]

Regression models provide quantitative metrics to determine model accuracy [81]:

- $Residuals\ (e) = Y_i - \hat{Y}_i$
- $e = obeserved - predicted$
- $MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$
- $RMSE = \sqrt{\frac{1}{n}\sum_{i=0}^{n} e_i^2}$
- $MAE = \frac{1}{n}\sum_{i=0}^{n}|e_i|$

Where:

- e = Residuals
- Yi = Observed Value
- $\hat{Y}_i$ = Predicted Value
- n = Number of Samples
- i = Sample number

2.5.2. Machine Learning Techniques Overview for this Study

This section reviews some of the essential ML techniques applicable to the field of study. The categories of machine learning and methods therein, are wide-ranging and are continually expanding. The selection of the applicable technique depends on the specific application, data available, and the required model output. Figure 2-19 gives an overview of the different ML categories and the relevant techniques employed within each category.

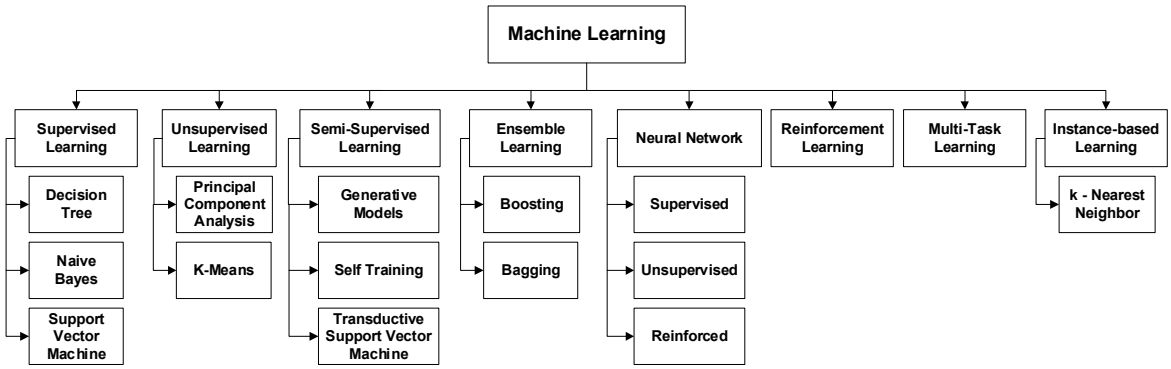

Figure 2-19 - Machine Learning Types - Source: Adapted from [78]

Datasets used for ML model development and testing requires a test and training dataset. The two datasets are a percentage of the original data set (example training set is 80%, and the test set 20%)[81].

Three approaches exist for the development of a PdM system, namely, data-driven, model-driven or a hybrid approach. The data-driven approach relies on collected sensor data for data analysis using data mining or ML. The model-driven approach uses an analytic representation of the system map out system behaviour [82].

Parameter estimation in PdM consists of either Cross-Sectional Forecasting or Time-Series Forecasting. The former referring to an estimate based on a specific condition where no measurements exist and the latter to predict the change over time. Essential data sources for a PdM model consists of fault history, maintenance and repair records and machine conditions. Typical ML algorithms for PdM consist of binary classification (probability of failure over time), regression models (calculate RUL of an asset), and multiclass-classification (determine the probability of failure in the future and assign a time interval for asset failure) [82].

The most important terminology used in predictive modelling includes [81]:

- Sample –Single, independent unit of data
- Training Set – Data used for modelling
- Test Set – Data used for testing the developed model
- Predictors – Input data for prediction
- Outcome – Output event predicted
- Continuous data – Data that is continuous over a numeric scale
- Categorical data – Data with specific values
- Sensitivity – Rate of correct event prediction for all samples including the event
- Specificity – Rate of prediction of non-events
- False Positive – False prediction of the positive class
- False Negative – False prediction of the negative class

The section below discusses the applicable ML techniques for this study.

2.5.2.1. Supervised Learning

Supervised learning consists of an ML algorithm that consists of labelled cases (form existing data), that can predict new cases. The output of the model will be a predefined

case. The algorithm can either consist of a classification or regression algorithm.[83]. The former being a categorical variable and the latter a quantitative output value [81].

Standard techniques for supervised learning includes:

2.5.2.1.1. Linear Regression

Regression analysis is a statistical technique to determine the relationship between variables. Scatterplots plot the relationship between variables based on the formula for LR below [84]:

$$\mathbf{y} = \boldsymbol{\beta_0} + \boldsymbol{\beta_1 x} + \boldsymbol{e} \qquad \textit{2.5 – Linear Regression Model}$$

Where:

- y = Numeric response
- $\boldsymbol{\beta}_0$ = Intercept
- $\boldsymbol{\beta}_1$ = Slope
- e = Statistical error to fit the slope

Regression models containing one regressor is referred to as simple linear regression models, whereas models containing multiple regressors is called multiple linear regression models[84]. The formula below expresses the mathematical relationship for the latter [81].

$$\mathbf{y_i} = \boldsymbol{b_0} + \boldsymbol{b_1 x_{i1}} + \boldsymbol{b_2 x_{i2}} + \cdots + \boldsymbol{b_p x_{ij}} + \boldsymbol{e_i} \qquad \textit{2.6 – Multiple Linear Regression Model}$$

Where:

- y_i = Numeric response for i^{th} sample
- b_0 = Intercept
- b_j = Estimated coefficient for the j^{th} predictor
- x_{ij} = Value of the j^{th} predictor for the i^{th} sample
- e_i = Random error of the model

LR model development includes three variables, namely, quantitative -, categorical - and indicator variables [84].

2.5.2.1.2. Decision Trees

A decision tree emulates a tree, which sorts attributes in groupings based on data values. Decision trees consist of nodes and branches as defined in the algorithm [83].

2.5.2.1.3. Naïve Bayes

The Naïve Bayes algorithm is based on the Bayesian probability theory and also allow for clustering or classification tasks. If the outcome of the algorithm is not specified, class assignment depends on the conditional probability of data values (typically used in unsupervised learning). If used in supervised learning, both target and outcome variables need to be specified to create Bayesian networks based on the conditional probability of the occurrence of the outcome [83].

2.5.2.1.4. Support Vector Machine (SVM)

The distance between two vectors on the same hyperplane is known as the margin, and the SVM algorithm draws boundaries or margins between classes. This algorithm maximizes the distance between each class and the nearest margin to reduce classification errors [83].

2.5.2.1.5. Logistic Regression

Logistic regression (LRR), a less complicated technique compared to SVM, is a classification algorithm that establishes linear classification boundaries. Youden's J Index expresses the prediction of false positives and false negatives [81]:

$$\boldsymbol{J = Sensitivy + Specificity - 1} \qquad \textit{2.7 – Youden's J Index}$$

2.5.2.1.6. k-Means Algorithms

The k-Means algorithms create groups of unlabelled data based on the mean distance between classes [77].

2.5.2.1.7. Random Forest

The Random Forest (RF) technique is a non-linear implementation of SVM and creates regression trees and often include bagging. Bagging techniques are low bias high variance techniques, which improves prediction accuracy [81].

2.5.2.2. Unsupervised Learning

Unsupervised learning consists of grouping data into clusters where no categorization exist. This model aims to learn new relationships within the data. Typical studies used boosting and bagging algorithms, as well as the k-Nearest Neighbours (k-NN) algorithm [83].

2.5.3. Section Summary

This section reviewed the data analytics evolution process, as well as ML algorithms used in PdM and CBM.

The next section reviews popular CBM and PdM approaches, as implemented in industry.

2.6. PdM and CBM Approaches

This section is a pre-cursor to case studies related to the study scope to understand the challenges experienced with previous PdM and CBM implementations.

As mentioned in the sections above, PdM aims to optimize equipment availability and reliability by prediction of future equipment states. CBM, however, focusses on equipment state monitoring to trigger alarms and prevent equipment failures through real-time monitoring of equipment [72].

Sakib and Wuest evaluated research related to PdM and CBM, and their findings are summarized below:

2.6.1. Markov Modelling

From the onset of PdM and CBM, prediction inaccuracies exist for equipment have multi-state failures. The hidden semi-Markov modelling approach determines the transition probabilities of equipment with multi-failure states. The piece-wise Markov modelling approach can be applied where random natural failures occur. These two Markov models enable the determination of machine states and the current number of jobs in the system [72].

2.6.2. Cost Maintenance

Establishing PdM systems in environments with high levels of uncertainty can come at a high cost. Reduction of the long-term mean maintenance cost is possible with the implementation of dynamic maintenance program predicting the RUL. Bayesian networks and Monte Carlo Simulation can model cost maintenance programs [72].

2.6.3. Scheduling

Scheduling maintenance activities can become cumbersome and the Hybrid multi-objective immune algorithm (H-MOIA) used in conjunction with least flexible job first (LFJ) and the longest processing time (LPT) algorithm, can be used for scheduling PdM activities to support the minimal impact of the disrupted operation on the schedule (MIDOS) system [72].

2.6.4. Bayesian Approach

PdM Bayesian approach can predict the equipment failure time and can schedule the required maintenance activity [72].

2.6.5. Neural Network Approach

Complex calculations and algorithms can be model using an artificial neural network (ANN) and determine the hazard rate and the MTBF based on real-time data [72].

2.6.6. Big Data Approach

With increasing data collection from various sources, the amount of data to be analysed for prediction is also growing. The RF algorithm can split large datasets into smaller sets to determine realistic outcomes. The use of an Autoregressive Moving Average (AMA) with Support Vector Regression (SVR), is used for unscheduled prediction of faults [72].

2.6.7. Time-to-Event Approach

Time-to-event estimation, also known as survival analysis, is used to determine a time to a specified event. This approach can also determine the RUL of equipment by combining the equipment life and CBM data [69]. Clark et al. suggest that survival analysis consists of two probabilities, namely survival and hazard. The former is the probability of surviving an event, while the latter is the instantaneous event rate for the

subject [85]. By combining maintenance, operating, and current health state data, prediction of the current equipment health state and future state is possible [69].

Determining the survival time of equipment is performed by evaluating a KM survival curve, that plots survival probabilities against time. The KM summaries also provide useful data of the survival function, in particular, the median survival time [85].

2.6.8. Section Summary

This section reviewed popular CBM, and PdM approaches and the next section evaluates case studies for different applications.

2.7. Case Studies Applicable to Study Scope

Due to limited literature available on PdM or CBM for CP systems, evaluation of case studies aims to inform the research design of this study.

2.7.1. Risk-Based PdM using Probabilistic Inference

Xu and Tang developed a PdM system based on risk classification and probabilistic inference for safety-critical systems in 2013. The main aim was to reduce equipment downtime and increase operational safety by managing the consequence of failure through a risk-based approach.

The design of the PdM algorithm consisted of system and sub-system identification, failure rate and pattern determination by using FMEA's, evaluating the risk using probabilistic inference, and determining the optimal maintenance strategy. For the algorithm, a 2-Step Temporal Bayesian Model (2-TBN) captured the system layout, and a conditional probability table (CPT) of the 2-TBNg based maintenance model captured the equipment failure data. 2-TBN models are typically used to model complex systems [73].

From the results obtained for an electrical track circuit case study, the authors were able to predict the required maintenance activity, equipment failure rates and an overall system optimal maintenance time based on the risk and system status. The algorithm, however, only predicted a system-wide optimal maintenance time and not per piece of equipment.

2.7.2. PdM Using A Multiple Classifier Approach

Susto et al. considered the use of a multiple classifier approach for a PdM system. This approach considered specifying numerous iterations of the maintenance cycle in the algorithm, except only the last occurrence, in an attempt to fine-tune the maintenance required by choosing broader failure horizons. The algorithm would thus run multiple times for each of the classifications specified [86].

The accuracy of the model output was verified with Monte Carlo Cross-Validation (MCCV) with semi-accurate results when compared to cross-validation. The misclassification calculation rate was also used as a performance metric to determine the accuracy of the classification [86].

The justification of the selection of the multiple classifier approach over other popular algorithms such as SVM or k-NN was due to the former's high computational requirements and the latter's low complexity non-parametric functionality. The case study presented by the authors, however, evaluated the model using both the SVM and k-NN algorithm [86].

A case study focusing on replacing tungsten filaments in an ion implantation facility presented good results that were usable by plant engineers by enabling multiple equipment health indicators and showed fair predictions on maintenance cost and activities. The SVM and k-NN algorithms also produced great prediction accuracy, although further implementation of Relevant Vector Machines (RVM) can improve the performance of the model [86].

2.7.3. Predictive Maintenance Architecture For Nuclear Infrastructure

Gohel et al. presented research findings on secure data transfer from field sensors and using the collected data to run ML algorithms for a PdM system at a nuclear facility. Secure data from the Internet of Things (IoT) devices was a shortfall in previous studies, and the model proposed by the authors ensures that the data cannot be accessed by third parties which can comprise the control system of the nuclear facility [87].

For the ML implementation, the authors used the Python scikit-learn software for modelling and testing the algorithm. The ML model consisted of both SVM and LRR algorithms. The former used to search for boundaries between features where class separation is possible and for prediction, the latter being used to describe outcomes of a single trial using multiple iterations. Performance evaluation of the two models consisted of calculating the difference between the predicted and actual class for specific inputs [87].

With the combination of the two algorithms, the authors tried to predict when an engine will fail within specified hourly cycles. For this prediction, the LR algorithm assigned two label values to the training dataset for each hourly cycle (either negative or positive). The label assignment was either "negative" if an engine will not fail within the next n cycles or "positive" if an engine will fail within the subsequent *n* cycles. The SVM algorithm determined the probability of an engine failure within a current hourly cycle using a scoring model where a lower score presented that the equipment was in a healthier state than others with higher scores (which indicates a failure) [87].

The ML framework predicted the required maintenance from real-time IoT sensor data with high accuracy and can extend to other applications (such as amperage spike detection, harmonic distortion detection and temperature increases) [87]. What was not apparent in the study was that the volume of data used for training and testing the algorithm and the SVM algorithm could present a high computation overload.

2.7.4. PdM in Industry 4.0 and Microsoft Azure

Paolanti et al. presented research findings by using the RF algorithm to predict the maintenance of electric motors and other equipment. Various sensors provided data to the ML model, and the analysis was done in the Microsoft Azure Cloud [82].

The authors ran various models to predict the spindle health status of a drive using built-in libraries from the Azure Machine Learning library. The ML model used a 70%:30% split for training and testing data sets (from sensor data) [82]. The results from the study were accurate, and the models were evaluated and tested in minimal time due to the existing libraries in Azure. No information was, however, available on how Azure combines algorithms and the source code to output a result. Another significance of this study was the rapid development of the ML framework using the Azure libraries.

2.7.5. MLP and SVM Algorithms for PdM of Centrifugal Pump

Orrù et al. presented an ML model for predicting centrifugal pump failures based on the SVM and Multilayer Perceptron (MLP) algorithms using the KNIME platform [88].

The authors defined two classification classes for determining the pump state (namely 1 for failure horizon and 0 for healthy horizon) using the SVM algorithm. The MLP algorithm implementation acted as an ANN for statistical analysis of non-linear data. The SVM algorithm presented a higher accuracy but a lower recall of positive cases, whereas the MLP algorithm presented a higher accuracy than the SVM algorithm [88].

The authors tried to implement two basic algorithms to indicate the simple process to set up a PdM system and achieved success, but future work requires improvement of availability metrics.

2.7.6. RUL Prediction

Ragab et al. predicted the RUL using a combination of equipment life data and the data received from the CM system. The modelling consisted of a combination of both a time-driven approach using the Kaplan-Meier (KM) technique and an event-driven technique to handle discrete operating data [69].

The authors estimated the performance of their model by comparing the actual RUL with the predicted RUL. A case study determined the RUL of a turbofan engine and assigned both a short-life (SL) and long-life (LL) state to the testing and training datasets using LAD [69].

The proposed technique does not require a threshold strategy to estimate the RUL and is not statically dependant and depends well, even with highly correlated covariate data [69].

2.7.7. Section Summary

This section reviewed case studies to extract design ideas from previous research.

2.8. Conclusion

The literature review covered topics applicable to the field of study, and the main objective was to gather the information that can aid in answering the research questions. Although there is not abundant literature available for predictive maintenance of CP systems, the literature review seeks to establish the relevant theoretical foundation to inform the research design.

The corrosion theory section provided an overview of how and why corrosion occurs and how to prevent it. A further in-depth overview of corrosion prevention, using CP systems, provided a necessary foundation that informs the research design of this study. Lastly, consultation of various industry standards regarding pipeline operations led to forming a basic framework that considered factors such as regulatory requirements, maintenance management, data management and operation of CP systems. The NACE SP0169-2013 standard defines the criteria for evaluating CP systems and is the baseline metric for this study.

A further literature review investigated reliability engineering principles that relate to the scope of this study, with a specific emphasis on maintenance strategies and reliability metrics. From this investigation, it was evident that CBM or PdM systems related to the scope of this study and formed the basis of the literature review for the next section, which evaluates the evolution of data analytics and ML techniques from relevant literature.

Based on the limited literature available for the scope of this study, further evaluation of PdM case studies provided general insight into typical techniques used for CBM and PdM system implementation (although not applicable to CP systems).

In conclusion, a reduction in the cost of corrosion of pipeline networks will sustain pipeline operations and ensure the asset is in use for its intended design life. The use of technology, such as remote monitoring and CBM and PdM systems, can aid to manage and maintain the primary equipment used in CP systems, namely the ICCP stations. The indirect impact of effective monitoring and maintenance of ICCP systems can lead to less corrosion exposure of TP's along the pipeline.

The next chapter focusses on the research design and methodology.

3. CHAPTER 3: RESEARCH DESIGN AND METHODOLOGY

3.1. Introduction

Kothari describes research as the “scientific and systematic search for pertinent information on a specific topic” [89].

Chapter 3 aims to establish the predictive maintenance framework design based on historical CP data evaluation for two selected pipeline sections. The literature review evaluated the standards and statutes that govern pipeline operations, and more specifically, the evaluation criteria for CP systems (NACE SP0169-2013). The CP criteria for instant-on establishes a CP pipe potential OW or state and is defined as the primary criteria for the modelling outcome of the predictive framework design. The statistical analysis results intend to evaluate the feasibility of the predictive maintenance framework (based on the prediction accuracy of each model). Furthermore, the ML techniques and case studies from previous work present various techniques for predictive maintenance frameworks in different industries.

The RBM and RCM principles are incorporated into the development of the maintenance matrix since the matrix will be based on conformance to the defined CP pipe potential OW. As mentioned in the literature review, the Kaplan-Meier Survival analysis or cycle time approach can be used to predict equipment failures and establishes a foundation for time suggestion of maintenance activities in this study.

The chapter layout is as follows [45]:

i. Research strategy – Describes the research approach and design methodology followed.
ii. Research context – Describes and justifies the selected research approach.
iii. Research data sources – Justifies the selection of data for the study.
iv. Data collection methods – Defines the data collection methods, strategy and instruments used.
v. Data analysis methods – Describes and justifies all techniques used for data analysis.
vi. Issues of trustworthiness – Describes the reliability, validity and reproducibility of the data
vii. Limitations and Delimitations – Describes the limitation and delimitations of the study.

3.2. Research Strategy

This study uses numerical data sets, and hence a quantitative study needs to be performed [90]. This quantitative study includes statistical treatment of the CP data sets to facilitate data analysis [89]. An empirical research design strategy enables the systematic evaluation of assumptions and the presented framework of this study.

The strengths of the empirical research design strategy include flexibility, outcome evaluation based on different research environments, control various variables to change the research outcome, improve analytical skills of the candidate and improves

internal validity [91],[92]. The main weakness of the empirical research design is the time required to perform the study and the availability of data.

To perform the quantitative research for this study, the candidate selected the empirical research design method to model, evaluate and assess the predictive maintenance framework based on two pipeline sections. The modelling will enable the candidate to identify, test and validate ICCP unit and TP states and the required maintenance output. The primary data source for this study is historical CP operating data collected from a CP SCADA system.

3.3. Research Context

The research context for this study consists of two pipeline sections that have either an FDU or TRU ICCP unit, to enable analysis and prediction based on the specific operating conditions, respectively.

For clarity, the pipe-to-soil potential (V_{CSE}), is also referred to as the CP pipe potential (V_{CSE}) and vice versa in chapters three to six. All CP pipe potentials are instant-on potentials.

3.3.1. Typical Pipeline CP System Design

A typical pipeline CP system consists of either a galvanic anode or ICCP rectifiers or a combination thereof, that provides CP current for a specific distance on the pipeline network [27]. As per the literature review, the design of the CP system depends on various factors and the as-built number of ICCP units can vary on the pipeline network. A pipeline network is a combination of transmission and numerous distribution pipelines. Figure 3-1 illustrates a typical high-level pipeline network which includes TRU's, an FDU and TP's:

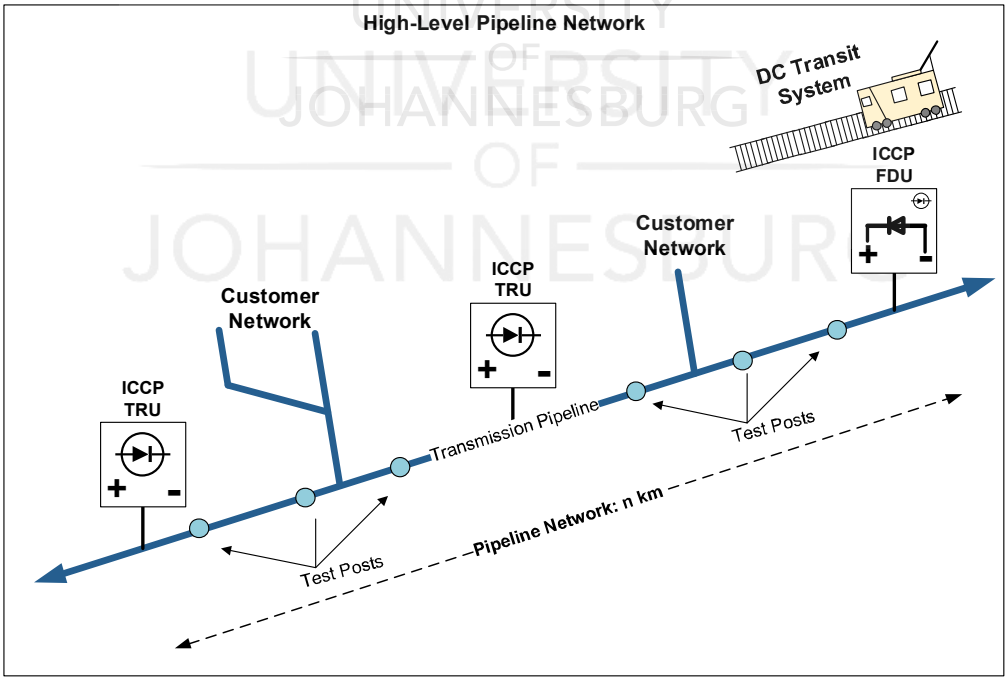

Figure 3-1 - Typical Pipeline Network CP System

CP systems can include other corrosion prevention equipment, such as NDU's, cross-bonds and ACM's, this study will, however, focus primarily on the data collected from ICCP units and a series of downstream TP's.

3.3.2. CP System Monitoring

Large pipeline operators place significant emphasis on remote monitoring of ICCP systems to determine the health of the ICCP systems. Typically, remote monitoring only reports the real-time process values and does not include data analytics which can potentially identify operational and failure patterns and hence aid in maintenance activities.

Remote monitoring of ICCP units typically enables data acquisition of the rectifier output voltage and current, as well as the CP pipe potential. Monitoring of digital signals can also aid to provide secondary status information of the ICCP unit operation.

3.3.3. CP System Effectiveness Evaluation

Real-time analysis of reported CP data typically consists of a value dead-band approach, which triggers an alarm if the reported value exceeds the defined operating band [54][93]. In areas with significant interference, an alarm can be triggered numerous time during a day, which can inform a reactive maintenance activity, but will also result in nuisance alarms. Applicable to this study, time filtering of dead-band alarms ensures it only triggers if the CP pipe potential is operating outside the dead-band for a certain amount of time.

The dead-band limits are typically the NACE SP0169-2013 criteria, consisting of a CP pipe potential between -850mV$_{CSE}$ and -1.1V$_{CSE}$ (dependant on the measurement type as defined in the standard) [11]. For pipelines in South Africa, the CP pipe potential dead-band is between -850mVCSE and -5.0VCSE due to the presence of significant stray current. In this study, the candidate also refers to this dead-band as the operating window (OW) or the protection band (PB).

Because CP pipe potentials usually vary at a low rate (due to polarisation), it is not always feasible to analyse raw data at very high sampling rates for ICCP units. Where significant stray current exists, the sampling rate can be adjusted to determine the magnitude of the interference problem.

3.3.4. CP System Effectiveness Challenges

Although the above alarming scheme can aid in informing the required maintenance activity, the execution of the maintenance can result in a high cost, if proper planning, scheduling and implementation are not considered for large pipeline networks. Furthermore, what is not apparent from the rectifier operation, is the effect of a low or high ICCP rectifier output at TP's along the pipeline.

Pipeline operators are only required to record the TP's once per calendar year [11][43], which can result in periods where a specific section of the pipeline has an increased corrosion risk, due to CP equipment damage or malfunction.

3.3.5. Factors Affecting Pipeline Maintenance

Maintenance and operation of existing pipeline networks can present additional challenges due to the following:

- Out-of-date master asset database, which does not include the commissioning data and as-built configuration of the pipeline and CP systems
- Missing operational and maintenance data
- Not extracting knowledge from CP data received

From a pipeline integrity perspective, the above can contribute to the following:

- Inability to accurately determine the pipeline integrity status
- Uncontrollable pipeline interference from stray current, DC transit systems, and foreign pipelines
- Escalating cost due to repetitive maintenance executed or only focussing on corrective maintenance
- Degrading pipeline safety

3.3.6. Research Focus Sections

This study will focus on two different pipeline sections, distinguished by the rectifier type installed, namely either a TRU or FDU. FDU's are usually installed next to DC transit systems and have a more significant impact if not operational. By splitting the analysis into two sections, a distinction is possible between the criticality of the equipment installed.

A typical rectifier installed in the industry has the following high-level circuit diagram:

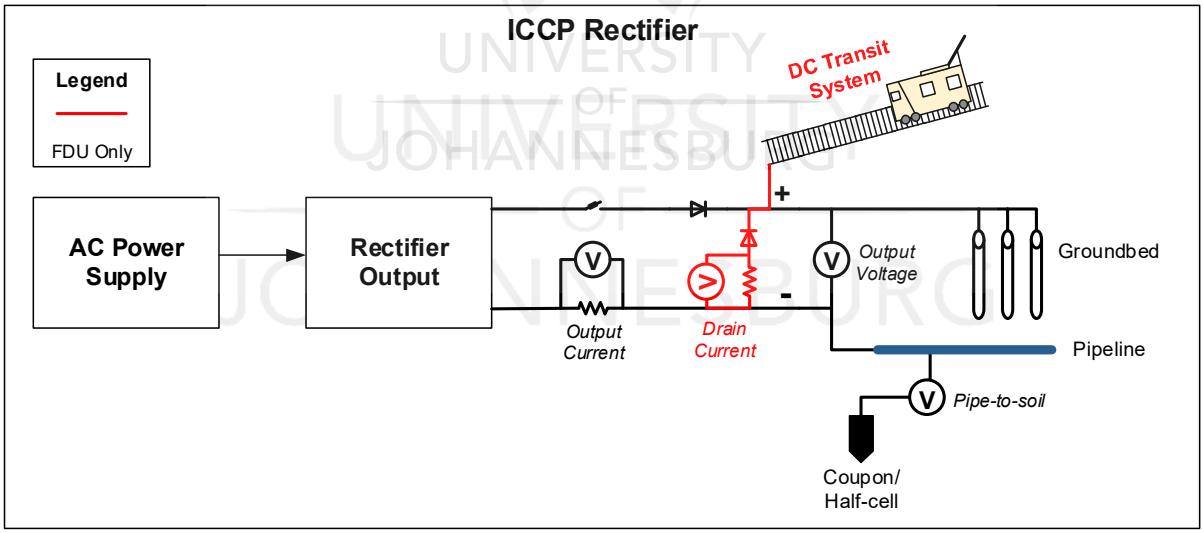

Figure 3-2 - ICCP Rectifier Wiring Diagram - Source: Adapted from [94]

The majority of CP rectifiers in South Africa uses a three-phase AC power supply. The rectifier converts the three-phase AC to pulsating DC using a silicon-controlled rectifier (SCR) stack or firing card. The rectifier output is adjustable to achieve a specific pipe-to-soil potential.

As indicated in the diagram above, the following continuous data points are available for this study:

- Rectifier Output Voltage
- Rectifier Output Current
- Rectifier Drainage Current (FDU only)
- Pipe-to-Soil Potential (V_{CSE})

All recordings are instant-on (when referring to the SP0169-2013 standard).

3.3.6.1. TRU Pipeline Section

As per the literature review, a TRU is an ICCP unit which supplies external current to a CP ground-bed to protect the pipeline against corrosion [28]. For this study, a 21km pipeline section, consisting of one ICCP TRU and 18 TP's, are selected to perform the data analysis.

Each of the 19 stations has continuous remote monitoring installed that sends data to a SCADA system which stores the data in a relational database. Periodic data is also available for the TP's.

The TRU pipeline section consists of the following:

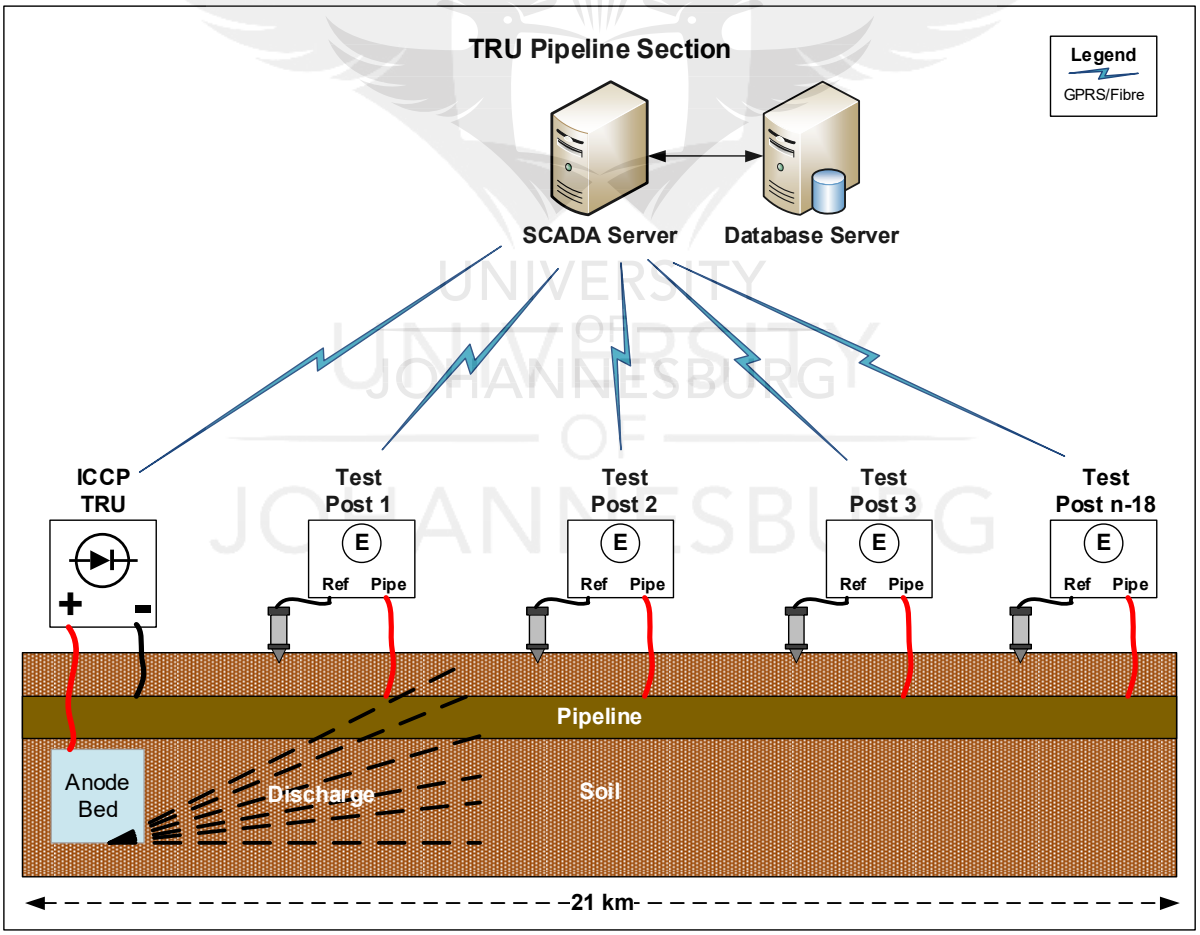

Figure 3-3 – TRU Pipeline Section - Source: Adapted from [28]

The above diagram depicts an ideal pipeline design where no stray current sources exist. The scope of this study will focus primarily on the data received from the remote monitoring system and does not include any known interference sources.

3.3.6.2. FDU Pipeline Section

As per the literature review, an FDU is an ICCP unit which supplies external current to a CP ground-bed to protect the pipeline against corrosion and additionally drains current pickup from DC transit systems back to the rail network [28]. The selection of an FDU section was motivated by the increased corrosion risk associated with a malfunctioning FDU [57].

For this study, an 8km FDU-protected pipeline section, consisting of one FDU and eight TP's, are selected to perform the data analysis. Each of the nine stations has continuous remote monitoring installed that sends data to a SCADA system which stores the data in a relational database. Periodic data is also available for the TP's.

The FDU pipeline section consists of the following:

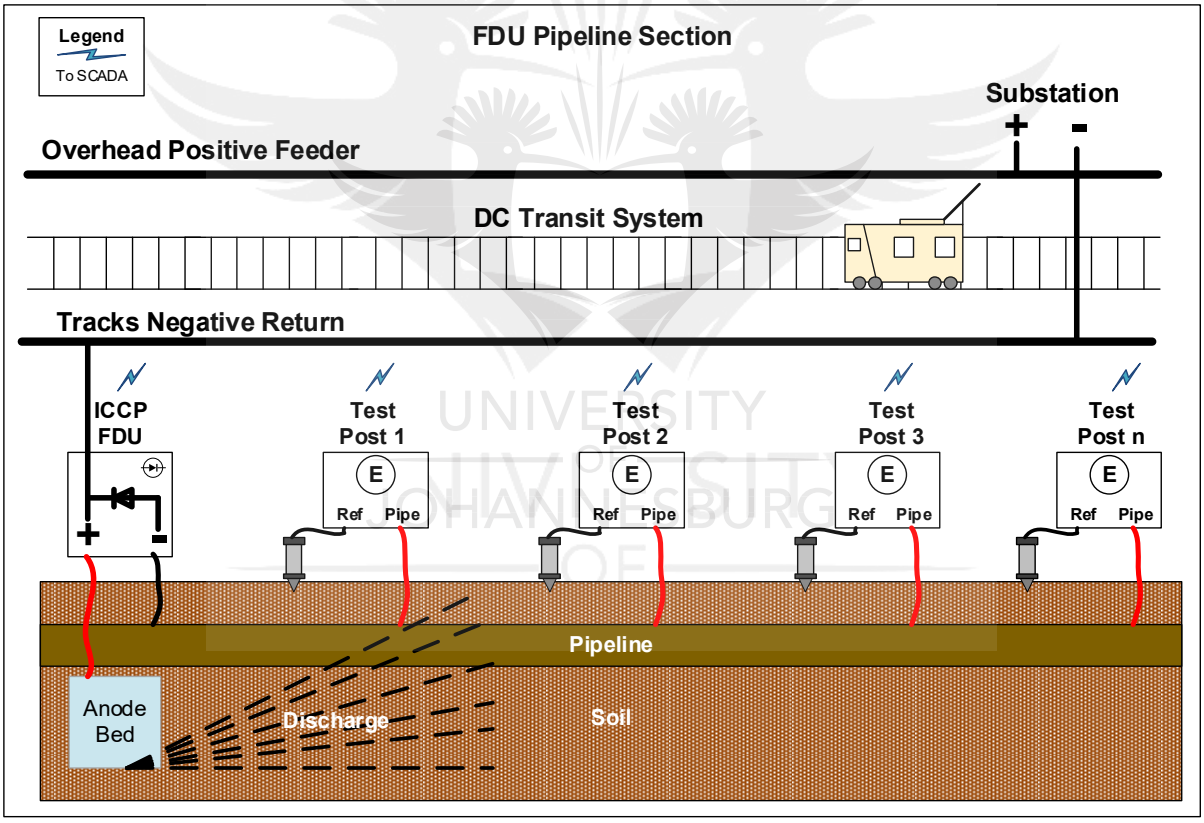

Figure 3-4 – FDU-Protected Pipeline Section - Source: Adapted from [28]

The above diagram depicts an ideal pipeline design where no stray current sources exist. The scope of this study will focus primarily on the data received from the remote monitoring system and does not include any known interference sources.

3.4. Research Sample And Data Collection

The data sets applicable to the research study objectives and research questions consists of raw data at different intervals collected from a CP SCADA system (for ICCP

units and data loggers), as well as manual field recordings. The exact data collected is discussed in the section below.

3.5. Data Collection Methods

Neumann suggests that for every study, data collection, from one or more sources, is required [95]. For this study, historical operating data from the selected ICCP units and TP's will enable the data analytics.

The data sources for this research consist of the following:

3.5.1. SCADA Process Data

The primary data for this study consists of historical process values from a SCADA system which monitors ICCP stations. The data points consist of the following:

- Rectifier Output Voltage
- Rectifier Output Current
- Rectifier Drainage Current (FDU only)
- Pipe-to-Soil Potential (V_{CSE})

All recordings are instant-on (when referring to the SP0169-2013 standard).

3.5.2. Logger Data

As an additional primary data source, the collection of TP measurement data enables the evaluation for pipe-to-soil measurements between ICCP units. The candidate collected this data from a historical recording database (Microsoft SQL). The data points consist of the following:

- Pipe-to-soil potentials
- Pipe AC potential

All recordings are instant-on (when referring to the SP0169-2013 standard).

3.5.3. Geographical Information System Data

As an additional secondary data source, asset location data was collected from a GIS system, to map GPS coordinates of TP's and ICCP units to the location of the actual recording. GPS coordinates enable visualization of the study results and to determine the distance between assets.

3.5.4. Datasheets and Manuals

Datasheets, manuals and working documents provide further information on actual equipment operation and maintenance required. Furthermore, consultation of the mentioned documentation can aid to remove uncertainty in the function or operation of a piece of equipment. A review of the statutory standards relating to pipeline operation and maintenance (for example, NACE and CFR standards) informed the research design.

3.5.5. Participation and Observation

The candidate was employed for the duration of this study and performed extensive system designs for telemetry systems and data analytics of SCADA systems. Furthermore, the candidate implemented various hardware and software solutions to monitor CP systems remotely.

3.6. Data Analysis Framework

Williamson et al. suggest that quantitative research consists of numerical data analysis through statistical techniques. Software tools exist for these tasks and include Microsoft Excel, Standard Analytical Software (SAS) packages such as R or Python, or IBM SPSS (Statistical Package for the Social Sciences) [96].

The R Studio IDE enabled data analysis and visualization for this study, and all CP data received was in a comma-separated values (CSV) file format.

The data analysis approach for this study consists of the following:

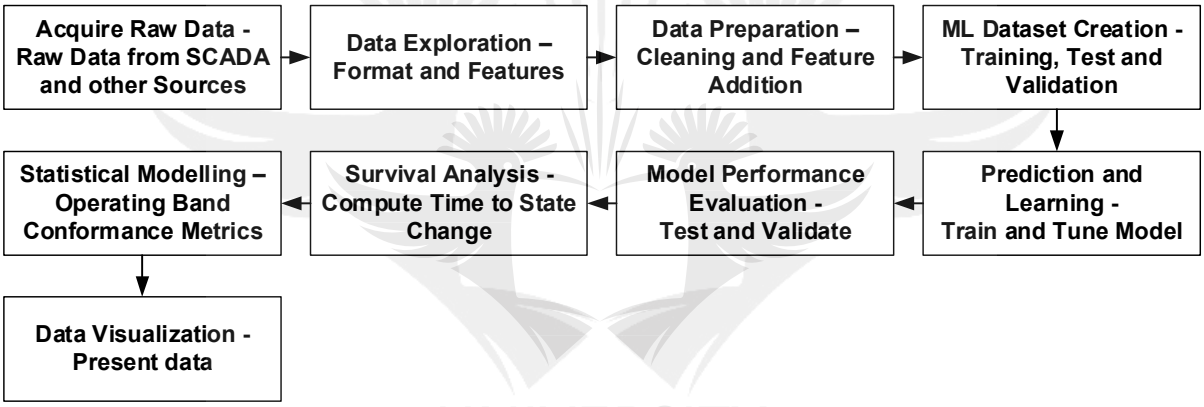

Figure 3-5 - Data Analysis Approach

This sections following describe the data analysis framework for this study.

3.6.1. Data Acquisition

As mentioned in section 3.5, data for this study consist of various sources. For the R analysis, historical data was retrieved from the SCADA system, logger database and manual recordings.

3.6.2. Data Exploration

Kuhn and Johnson classify data in two formats, namely continuous or categorical. The former consisting of numerical values only and the latter being discrete [81].

A first inspection of the raw SCADA data for ICCP units includes the following columns:

Date	Time	I1.value	X	I2.value	X.1	T1.value	X.2	V1.value	X.3	V2.value	X.4	X.5
2/1/2020	02:00:00	19.85		0.95		28.4	*	24.73		-9.34		NA
2/1/2020	02:00:30	19.85		0.95		28.4	*	24.73		-9.34		NA
2/1/2020	02:01:00	19.82		0.95		28.4	*	23.3		-9.02		NA

Table 3-1 - Initial Glance - ICCP Data

The data for this study is continuous and follows in a sequence based on the time and date columns. Anderson suggests that data which follows an ordered sequence with n observations is known as time-series data [97].

3.6.3. Data Preparation

Data preparation is required to ensure the data for this study is in the correct format and error-free [96].

3.6.3.1. Data Transformation

Williamson et al. suggest the transformation of data to make it more suitable for analysis [96].

The sections below describe the data transformation for this study.

3.6.3.1.1. Data Cleaning

From table 9 above, various columns exist with no data or string values in a regular numeric column. Removing columns with no data aims to reduce the size of the dataset. Column formatting, such as setting the data types and column name to a more user-friendly name, improves the model accuracy and assists in the coding process.

The column names are as follows:

- Date = Date of measurement
- Time = Time of measurement
- IOut = Rectifier Output Current
- IDrain = Rectifier Drainage Current (FDU only)
- VOut = Rectifier Output Voltage
- VCSE = Pipe-to-soil potential

Rows with erroneous values from the SCADA system (example, when the SCADA tag status is bad) or where R detected fields with no data (NA values) can result in a coding exception and inaccurate results. The candidate deleted all the erroneous rows, and the relevant dataset columns are shown below:

Date	Time	IOut	IDrain	VOut	VCSE
2/1/2020	02:00:00	19.9	0.95	24.7	-9.34
2/1/2020	02:00:30	19.9	0.95	24.7	-9.34
2/1/2020	02:01:00	19.8	0.95	23.3	-9.02
2/1/2020	02:01:30	19.9	0.95	23.4	-8.96
2/1/2020	02:02:00	19.8	1.52	23.5	-8.96

Table 3-2 - ICCP Data Cleaning

3.6.3.2. Feature Engineering

The sections below discuss the feature engineering tasks applicable to this study.

3.6.3.2.1. Timestamp

The first transformation of the dataset is to convert the date column into the format "yyyy/mm/dd". This format is required in R to calculate the difference between two date ranges.

The dedicated date and time columns increase the coding complexity in R, and a new column was created, called "Timestamp", which concatenates the Date and Time column to provide value in the format: "yyyy/mm/dd hh:mm:ss".

3.6.3.2.2. Index Column

An ID column, with an increasing numeric value for each row, was created to enable filtering by a numeric index number. The ID column value of each row increments by a numeric value of one.

3.6.3.2.3. Unit Type Column

A unit type column was created to indicate if the data originated from a TP, TRU, or an FDU:

1. TP
2. TRU
3. FDU

3.6.3.2.4. Event Time and Cumulative Time

For this study, the time between the status column events is required to calculate time statistics and to perform time-to-event analysis. For this analysis, two columns were created to calculate the duration between status events as well as the cumulative time per status.

3.6.3.2.5. Status and StatusNum Column

Part of the scope of this study is determining the health state of an ICCP unit, and for this indication, a health status indicator is required. The health status indicator does not exist in the initial data set and is defined as through a combination of statistical process control (SPC) control charts [98] and the NACE SP0169-2013 criteria for instant-ON potentials [11].

An operating window (OW) or protection band (PB) needs to be defined for the pipe-to-soil potential. For this study, the NACE SP169-2013 criteria were used as the maximum value, namely -0.85V_{CSE} and the minimum value was selected as -5.0V_{CSE}. Status label values were defined as follows:

Status Label Definition			
Label	**Value**	**Definition**	**Criteria**
P	1	Pipeline is protected	CP pipe potential is within the window of -0.85VCSE and -5.00VCSE
OP	2	Pipeline is over-protected	CP pipe potential more electro-negative than -5.00VCSE
UP	3	Pipeline is under-protected	CP pipe potential more electro-positive than -0.85VCSE

Table 3-3 - Status Label Definition

The status column is a categorical variable in this study, while the StatusNum column is a numeric representation of the status label (value from 1 to 3).

3.6.3.2.6. Rectifier Operational Column

A column was created to determine if the rectifier is operational based on the output voltage and current values received as well as the potential pipe state (P). Some rectifiers might be operating in a mode which might switch off the output current for a specific period, or an instrument error can be present [40], [44]. Rectifier mode data was not available for this analysis, and the rectifier operational status is determined using the criteria below:

Rectifier Operational Definition				
Column Name	**Column Value Range**	**Criteria 1**	**Criteria 2**	**Criteria 3**
RectOper	**0..3**	Vout > 0.0 & Iout > 0.0	Vout > 0.0 & Iout = 0.0 & Status = "P"	Vout = 0.0 & Iout > 0.0 & Status = "P"

Table 3-4 - Rectifier Operational Definition

3.6.3.2.7. CP Current Spread Factor on Pipeline Section

During the design phase of a CP system, the CP current spread from a ground bed, or sacrificial anode is calculated for a specific distance. This information was not available for this study, and a current spread factor was assumed based on the TP distance from the rectifier.

3.6.3.2.8. Rectifier Risk Level

A risk rating was assigned for data received from the different CP equipment (based on the criticality of the equipment):

- TP = 1
- TRU = 2
- FDU = 3

3.6.3.2.9. CP pipe potential Risk Columns

Evaluating the risk of exceeding the CP pipe potential OW, a pipe-potential risk-factor was created, consisting of three columns that classify CP pipe potentials in terms of their risk of exceeding their OW:

Risk Column Definition					
Column Name	**Column Value Range**	**V_{CSE} Range for 1**	**V_{CSE} Range for 2**	**V_{CSE} Range for 3**	**V_{CSE} Range for 4**
RiskLevelOP	**0..4**	-5.00V_{CSE} to -7.00V_{CSE}	-10.00V_{CSE} to -7.00V_{CSE}	-15.00V_{CSE} to -10.00V_{CSE}	< -15.00V_{CSE}
RiskLevelUP	**0..4**	-0.85V_{CSE} to 0.00V_{CSE}	0.00V_{CSE} to +1.50V_{CSE}	+1.50V_{CSE} to +3.50V_{CSE}	> +3.50V_{CSE}
RiskLevelP	**0..1**	-4.99V_{CSE} to -4.7V_{CSE} OR -1.00V_{CSE} to -0.86V_{CSE}			

Table 3-5 - Risk Column Definition for CP pipe potential

3.6.3.2.10. Stray Current Risk Column

The presence of stray current was evaluated based on the voltage range between two sequential records:

Risk Column Definition				
Column Name	**Column Value Range**	**Var_{CSE} Range for 1**	**Var_{CSE} Range for 2**	**Var_{CSE} Range for 3**
Stray Current	**0..3**	7 < Var_{cse} >= 5	10 < Var_{cse} >= 7	Var_{cse} >= 10

Table 3-6 - Risk Column Definition for Stray Current

3.6.3.2.11. CP Health Indicator

The CP health indicator for a pipeline section was created for this study that considers the following: the supplying rectifier status, the unit risk level, the CP pipe potential OW risk, and current spreading factor. The CP health indicator is calculated per piece of equipment and averaged out to determine the CP system health for the pipeline section:

$$\boldsymbol{R_{CPU}} = ((\boldsymbol{U_T} + \boldsymbol{R_U}) - \frac{\boldsymbol{U_T}+\boldsymbol{R_U}}{\boldsymbol{U_O}} + (\boldsymbol{F_{OP}} \times \boldsymbol{R_{OP}}) + (\boldsymbol{F_{UP}} \times \boldsymbol{R_{UP}}) + (\boldsymbol{F_P} \times \boldsymbol{R_P})) \times \left(\mathbf{1} + \frac{\boldsymbol{D_U}}{\boldsymbol{D_T}}\right)$$

3.1 – CP Unit Risk Indicator

Where:

- R_{CPU} = CP Risk Indicator
- U_T = Unit type (FDU, TRU or TP)
- R_U = Unit risk level (FDU, TRU or TP)
- U_O = Unit operational
- R_{OP} = Risk level of CP pipe potential (over-protection)
- R_{UP} = Risk level of CP pipe potential (under-protection)
- R_P = Risk level of CP pipe potential (protected)
- F_{OP} = Constant risk factor of CP pipe potential (over-protection)
- F_{UP} = Constant risk factor of CP pipe potential (under-protection)
- F_P = Constant risk factor of CP pipe potential (protected)
- D_U = Unit distance on the pipeline
- D_T = Total pipeline distance

If the effect of interference decays along the pipeline, the risk indicator formula is:

$$\boldsymbol{R_{CPU}} = ((\boldsymbol{U_T} + \boldsymbol{R_U}) - \frac{\boldsymbol{U_T}+\boldsymbol{R_U}}{\boldsymbol{U_O}} + (\boldsymbol{F_{OP}} \times \boldsymbol{R_{OP}}) + (\boldsymbol{F_{UP}} \times \boldsymbol{R_{UP}}) + (\boldsymbol{F_P} \times \boldsymbol{R_P})) \times \left(\frac{\mathbf{1}}{\mathbf{1}+\frac{\boldsymbol{D_U}}{\boldsymbol{D_T}}}\right)$$

3.2 – CP Unit Risk Indicator – Inverse Effect

Where:

- R_{CPU} = CP Risk Indicator
- U_T = Unit type (FDU, TRU or TP)
- R_U = Unit risk level (FDU, TRU or TP)

- U_O = Unit operational
- R_{OP} = Risk level of CP pipe potential (over-protection)
- R_{UP} = Risk level of CP pipe potential (under-protection)
- R_P = Risk level of CP pipe potential (protected)
- F_{OP} = Constant risk factor of CP pipe potential (over-protection)
- F_{UP} = Constant risk factor of CP pipe potential (under-protection)
- F_P = Constant risk factor of CP pipe potential (protected)
- D_U = Unit distance on the pipeline
- D_T = Total pipeline distance

The CP unit health is calculated as follows:

$$\boldsymbol{H_{CPU}} = \left(\mathbf{1} - \frac{\boldsymbol{R_{CPU}}}{\boldsymbol{T_{CPU}}}\right) \times \mathbf{100}\% \qquad \text{3.3 – CP Unit Health Indicator}$$

Where:

- H_{CPU} = CP unit health Indicator
- R_{CPU} = CP Risk Indicator
- T_{CPU} = Total nr of CP units on the pipeline

The overall health indicator per pipeline section is as follows:

$$\boldsymbol{H_{CPO}} = \frac{\boldsymbol{H_{CPU1}+H_{CPU2}++H_{CPUn}}}{\boldsymbol{n}} \qquad \textit{3.4 – Overall Pipeline Section CP Health Indicator}$$

Where:

- H_{CPO} = Overall Pipeline Section CP Health Indicator
- H_{CPU} = Overall Pipeline Section CP Health Indicator
- n = Number of units on the pipeline

A baseline constant risk factor was initially selected:

- Under-protected is 2.7 due to the most significant impact on the pipeline
- Over-protected is 2.0 due to a less significant impact when compared to under-protection
- Protected is 1.15 as this is the desired state of the pipeline

The factors above were adjusted in the evaluation of both an FDU or TRU.

Appendix B contains the different health results based on the above CP Unit Health formula for the three units type (FDU, TRU or TP).

3.6.3.3. Feature Selection

Since that data for this study is collected from a variety of sensors, feature selection or development will speed up computation and increase model accuracy [99]. The following columns were selected for predictive modelling:

- Rectifier Process Variables (IOut, VOut and IDrain) for use as predictors and logical programming inputs.

- CP pipe potential (Vcse) for use as an outcome variable and to determine prediction accuracy. They are also used as a variable for descriptive statistics.
- All columns discussed in the feature engineering section.

3.6.4. Machine Learning Model

Kuhn and Johnson suggest that prediction of future trends, with a specific probability, is possible by using historical data. Ayres further suggests that any prediction is not meant to replace human intuition, but rather complement it [81].

The most important aspect of this study is to determine if a predictive modelling or an extended CM approach is feasible based on the datasets collected for the pipeline CP systems. This section discusses the three main sections of the ML model, namely the classification algorithm for equipment state prediction and the survival analysis to predict time-to-state change.

The ML model for this study includes the following steps:

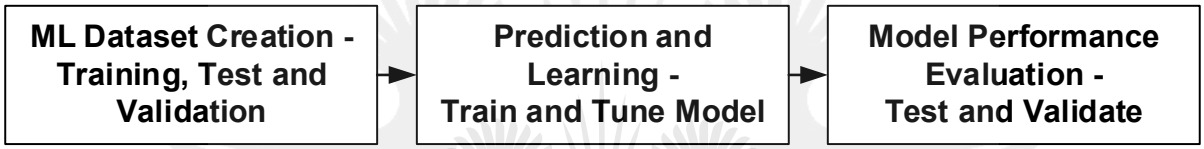

Figure 3-6 - ML Model Steps

The ML model steps are discussed in the sections below.

3.6.4.1. Training, Test and Validation Datasets

Saxena and Saad suggest using three different sizes of the original dataset for an ML project, namely a training, test and validation data set [99]. The training data set is used only for modelling, while the test and validation sets are used for model performance [81]. For this study, the training and test datasets were created based on scenario-specific ratios and are discussed in the next two chapters.

3.6.4.2. Prediction and Learning

The prediction and learning process consists of model development, testing and evaluation and model selection [81]. This section consists of the prediction of pipe-to-soil potentials, the equipment state and the suggested maintenance activity.

3.6.4.2.1. Pipe-to-Soil Potential Prediction

The first step in the modelling was to predict the CP pipe potential using a variety of ML techniques (without tuning) using the caret package in R [15].

3.6.4.2.2. Equipment State Prediction

Equipment health prediction is a crucial feature of a predictive maintenance system [100]. The equipment health prediction principle was extended for this study to consider whether an ICCP unit is operating at either of the three defined states, namely OP, P and P. A classification approach was adopted and included various combinations of predictors to improve the prediction accuracy. Different ML models

were evaluated in this study and will be discussed in chapter four and five. For the classification model, the caret package was used in R[15].

3.6.4.2.3. Maintenance Activity Suggestion

Based on the equipment state prediction, a maintenance activity was suggested to remedy the indicated fault. Based on the literature review and consultation with industry experts, the following maintenance matrix was defined:

Maintenance Activities				
Index	**Fault**	**Overall Risk Level**	**Time Limit (Hours)**	**Required Remedial Action**
1	Over-protection	1	4	Monitor and if required, adjust supplying rectifier
2	Over-protection	2	8	Adjust supplying rectifier output and monitor performance
3	Over-protection	3	16	Adjust supplying rectifier output and monitor performance
4	Over-protection	4	24	Adjust supplying rectifier output, investigate possible interference and monitor
5	Under-protection	1	4	Monitor and if required, adjust supplying rectifier
6	Under-protection	2	8	Adjust supplying rectifier output and monitor performance
7	Under-protection	3	16	Adjust supplying rectifier output and monitor performance
8	Under-protection	4	24	Adjust supplying rectifier output, investigate possible interference and monitor
9	Stray Current Interference	1	N/A	Investigate causes and adjust rectifier. Initiate interference mitigation projects
10	Rectifier not supplying current	1	8	Investigate rectifier and resolve the issue
11	Rectifier not draining current	1	8	Investigate rectifier and resolve the issue

Table 3-7 - Suggested Maintenance Matrix

The maintenance activities include both consideration for the risk level and the time aspect of the data evaluated. The maintenance activity was modelled using a classification ML model. The time component of the maintenance activity is covered in the survival analysis section of this chapter.

3.6.4.3. Model Performance Evaluation

Kuhn and Johnson suggest two model performance evaluation metrics to determine the model accuracy. These metrics include plotting actual values against predicted values or calculation of the model RMSE [81]. Chai and Draxler suggest the MAE can also be used for model evaluation and reasons that one is not superior over the other [101]. The RMSE is the square root of the average prediction square error value for the dataset while the MAE is the mean absolute distance between two points on a vertical or horizontal plane.

The RMSE formula is [101]:

$$\boldsymbol{RMSE} = \sqrt{\frac{\mathbf{1}}{\boldsymbol{n}}\sum_{\boldsymbol{i}=\mathbf{0}}^{\boldsymbol{n}} \boldsymbol{e_i^2}} \quad \textit{3.5 – Root Mean Squared Error}$$

Where:

- RMSE = Root Mean Squared Error
- n = Nr of samples
- e = Error between the predicted and actual value

The MAE formula is [101]:

$$\boldsymbol{MAE} = \frac{\mathbf{1}}{\boldsymbol{n}}\sum_{\boldsymbol{i}=\mathbf{0}}^{\boldsymbol{n}} |\boldsymbol{e_i}| \quad \textit{3.6 – Mean Absolute Error}$$

Where:

- MAE = Mean Absolute Error
- n = Nr of samples
- e = Absolute error between the predicted and actual value

3.6.4.4. Survival Analysis

Survival analysis is concerned with the time to an event under question [85]. Relevant to this study is the estimation of the time to a state change of either P, OP or UP. The estimated time can also be used for maintenance time estimation. For this estimation, an analysis needs to be performed that predicts the probability and survival time of a specific state. Survival analysis was performed using the event time and accumulated time column values as discussed in the feature engineering section.

3.6.4.4.1. Time-to-State Analysis

The time-to-state analysis was modelled using two functions, the Kaplan-Meier (KM) models available in the Survival package in R and assigning cycle times for three defined states [102].

Included in the analysis, was the conversion of the data set to a timestamp and status ordered data set with a cumulative time event calculation.

The survival function used in this study is based on the survival analysis of cancer patients according to a trial study performed by Clark et al. [85]:

$$S(\boldsymbol{t_j}) = S(\boldsymbol{t_{j-1}})\left(\mathbf{1} - \frac{\boldsymbol{d_j}}{\boldsymbol{n_j}}\right) \quad \textit{3.7 – Survival Function}$$

Where:

- $S(t_j)$ = Probability of being alive at time t_{j-1}
- n_j = Nr of samples alive before time t_j
- d_j = Nr of events at time t_j

To enable a running estimation of state change, the following cycle times were assigned to the two problematic states, that decrements as new values are received that matches the state:

- OP = 40 Hours
- UP = 24 Hours

This approach enables the prediction of time-to-state, as well as the actual time when the cycle has ended. These results can inform the maintenance schedule. The trend component of a decomposed time-series object will also be evaluated to determine the maintenance based on different evaluation periods.

3.6.4.4.2. Time-to-State Analysis Performance Evaluation

Evaluation of survival time model performance consists of an evaluation of both the KM survival curve and the survival summary data from the survival package. The trend components were visually analysed using a line graph, and the cycle time was estimated in Microsoft Excel.

3.6.5. Descriptive Statistics – Operating Band Conformance Metrics

The following statistical calculations enable further evaluation of the CP effectiveness based on a specific time window (example every 6 hours):

- Minimum, maximum and average
- Standard deviation and average
- Reliability KPI's - Conformance to OW (percentage and time statistics)
- Presence of stray current

3.6.6. Data Visualization

The results of the study objects were displayed using graphs, tables, screenshots and Microsoft Visio.

3.7. Ethical Considerations

No ethical considerations exist for this study.

3.8. Reliability, Validity and Reproducibility

Thomas suggests that the reliability of any study depends on the reproducibility of the research results on different occasions [103]. As this study is conducted using specific datasets collected from a SCADA system and study-specific coding in R, the reliability depends on two factors, namely the R source code and the actual data itself. The former refers to reproducing the study with the actual source code used in this study and the latter to the actual datasets used.

The validity of this study is based on the predictive modelling results, as described in section 3.6.4.3.

3.9. Study Limitations and Delimitations

Theofanidis and Fountouki define study limitations as any weaknesses that can affect the study that is not under the control of the researcher. In contrast, delimitations are boundaries set by the researcher to narrow the scope of the study to ensure the research objectives are met [104].

The limitations of this study are listed below:

- The accuracy of the raw data received from the SCADA and logger database due to faulty instruments, incorrect installation, incorrect polarity or using non-calibrated instruments.
- Incorrect asset location information received that will affect the ML model output.

The delimitations of this study are listed below:

- The scope for the data analysis consists of two sections, namely the FDU and TRU pipeline sections. Due to the length of pipeline networks, this boundary ensures that the research scope does not increase, and data analysis becomes impossible.

3.10. Conclusion

This chapter discussed the research design and methodology related to the scope of this study. From the available literature, the most suitable approach for this study was the empirical research design approach based on predictive modelling design framework suggested for this study.

The research context section discussed the NACE SP0169-2013 criteria to determine the effectiveness of the CP system. This study's research focus areas were also split into two sections; namely, the FDU and TRU protected pipelines, to enable data evaluation based on the criticality of the ICCP units and different operating modes.

Based on the research context and objectives, historical CP data was required for this study and was collected from four sources. The primary source being the SCADA system data and the CP logger recordings to feed data for the analysis. Data was also collected from a GIS system to determine the distance between ICCP units and TP's. Relevant datasheets, manual and standards were also consulted to ensure that the analysis conforms to the as-built design and the current statutory requirements. Lastly, a consultation was performed with industry experts where additional information regarding the operation and maintenance of CP systems was required.

The data analysis section discussed the various activities to ensure that the data is in a format that enables analysis. Activities included formatting rows and columns to ensure the data set is small and does not contain corrupt data. Additional features were added to the dataset that will enhance the outcome of this study, namely the status labels, risk determination, and event times. The ML model steps were discussed in the modelling process and consist of creating the training and test datasets, learning and predicting, CP pipe potential prediction and state prediction, maintenance activity

suggestion, as well as the survival analysis using KM and cycle times. The statistics computed for the datasets are also mentioned as well as the data visualization of the results obtained in R.

Lastly, this chapter discusses the ethical considerations, reliability and validity of the study and the study limitations and delimitations. The reliability and validity of the analysis depend on the availability of the R source code and the datasets used. The study limitations include the issue of CP data accuracy, whereas the delimitations considered narrow the scope for the study to only two pipeline sections.

The next chapter will discuss the data exploration results in an attempt to answer the stated research questions.

4. CHAPTER 4: EXPLORATORY DATA ANALYSIS

4.1. Introduction

The focus of this chapter is to perform an exploratory data analysis on the datasets received for this study to inform the prediction and learning phase of the ML modelling process.

This chapter consists of the following sections:

i. Evaluation of software packages to use for the ML modelling
ii. Exploring the datasets for the two focus sections of this study
iii. Exploring the descriptive statistics defined in chapter three
iv. Long-term time-series analysis with relevant use-cases

4.2. Evaluation of Software Tools for Statistical Analysis

Various software tools exist for statistical data analysis, whether open-source or paid versions. The three most popular tools are either Python, R or SAS [105].

Brittain et al. performed research to review the performance of both Python, R and SAS. The results indicated that one does not necessarily perform better than the other, and selection of either of the three packages depend on the computer resources, coding complexity and analysis objectives [105]. The table below summarizes the various qualitative attributes of the three tools.

Qualitative Attributes Comparison			
Attribute	**Python**	**R**	**SAS**
Packages Available	133915	10000	Integrated
Data Handling	RAM	RAM	Hard drive
Online Error	Yes	Yes	Yes
Numbers and Text	Yes	Yes	Yes
Interactive and Programmed	CLI & IDE	CLI & IDE	CLI & IDE
Complex Data Structures	Yes	Yes	Yes
Missing Values	Yes	Yes	Yes
Linear Algebra	Yes	Yes	Yes
Graphics	Yes	Yes	Yes

Table 4-1 - Qualitative Attributes for Software Tools [105]

As mentioned in chapter three, the candidate selected the R software tool for data analysis in this study, and the coding in R consists of various packages as referenced throughout this document. The R Studio IDE provides both a coding and visualization interface.

4.3. Dataset Exploration

The first section starts with an exploration of the raw data for both a TRU and FDU, respectively, to explore operational patterns, the next section focusses on the two

pipeline sections defined in chapter three and the last section investigates time-series analysis applicable to this study.

4.3.1. TRU Data

4.3.1.1. Overview

Figure 4-1 represents a wiring diagram of a typical TRU, with the measurement points indicated in red:

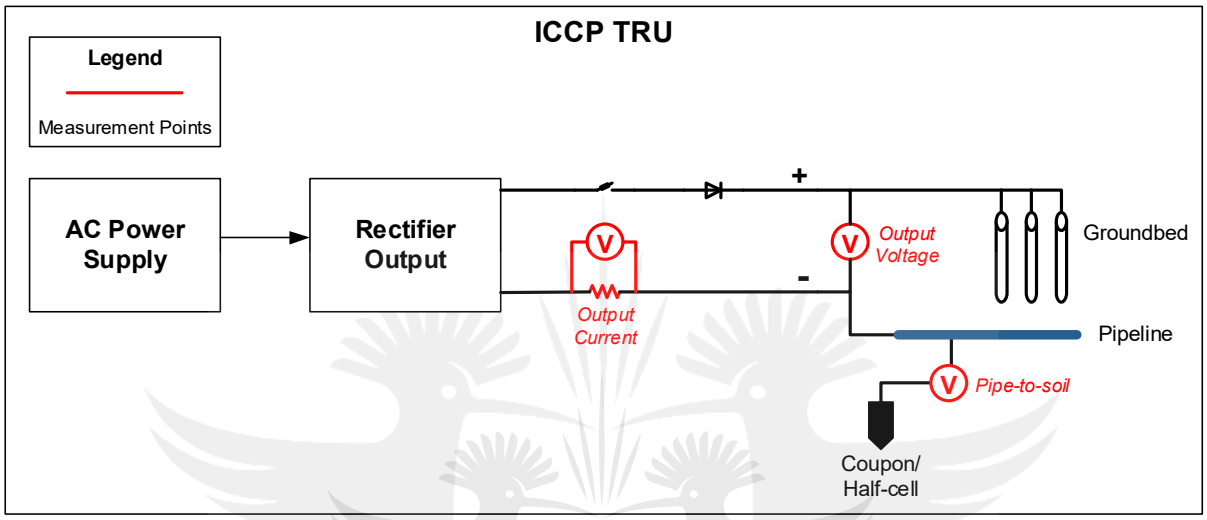

Figure 4-1- TRU Wiring Diagram - Source: Adapted from [94]

Continuous data from a TRU received for this study consists of the following measurement points or process values (PV):

- Rectifier output voltage
- Rectifier output current (V_{DC})
- Pipe-to-soil potential (V_{CSE})

Ohm's law governs the rectifier output voltage and current, whereby the external resistance determines the relevant voltage and current values. The pipe-to-soil potential is affected by the amount of current, either in a positive direction (decrease in CP current) or negative direction (increase in CP current). This statement, however, assumes no stray current is present.

When referring to the NACE SP0169-2013 standard for the criteria for CP protection, the measurement method needs to be recorded [11], [46]. For all datasets in this study, no known IR drop exists, and the CP pipe potentials (V_{CSE}) are instant-on. A constant IR-drop factor can adjust the CP pipe potential to compensate for the IR-drop; however, no IR-drop compensation took place in this study (future work).

4.3.1.2. Dataset Columns

Based on the data cleaning and feature engineering activities mentioned in chapter three, the dataset columns are as follows:

```
        Date     Time  IOut IDrain  VOut  Vcse           Timestamp ID Status StatusNum UnitType RectOper
1 2020-06-01 02:00:00 19.89   0.76 18.01 -8.42 2020-06-01 02:00:00  1     OP         2        3        1
  RiskFactor RiskLevelOP RiskLevelUP RiskLevelP OPFactor UPFactor PFactor DistanceFactor TotalDistance
1          3           3          0          0      2.5      3.1       1              1             8
  TotalRiskPos TimeFactor IntervalNr CPRiskLevel CPHealth
1           19       0.15          1         7.5 60.52632
```

Figure 4-2 - ICCP TRU Dataset with Indicators

4.3.1.3. Steady-state PV's

Upon examining the raw data from the various TRU's, the steady-state PV's are theoretically supposed to vary within a defined control band (as set on the rectifier). The absence of spikes (positive or negative), indicates that no stray current is present.

4.3.1.3.1. All TRU PV's

The graph below illustrates the rectifier output voltage and current, as well as the instant-on CP pipe potentials (V_{CSE}) from a TRU regulating the CP pipe potential at approximately -3.0V_{CSE}:

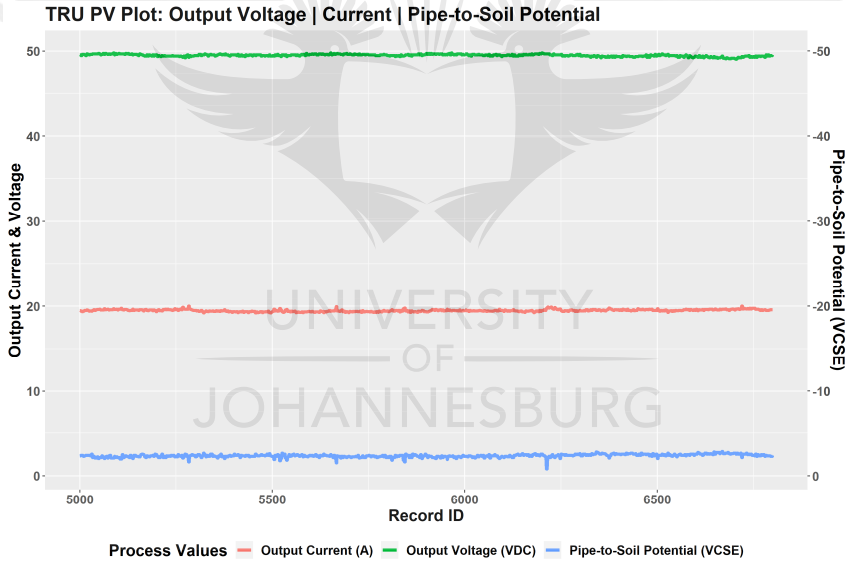

Figure 4-3 - TRU PV Line Graph (All Measuring Points) – Regulating at -3.0V_{CSE}

4.3.1.3.2. Pipe-to-Soil Potentials – Regulating within OP

Figure 4-4 illustrates the instant-on CP pipe potentials (V_{CSE}) from a TRU regulating the CP pipe potential within the defined OW. The minimum and maximum OW setpoints (SP) acts as a visual OW guide.

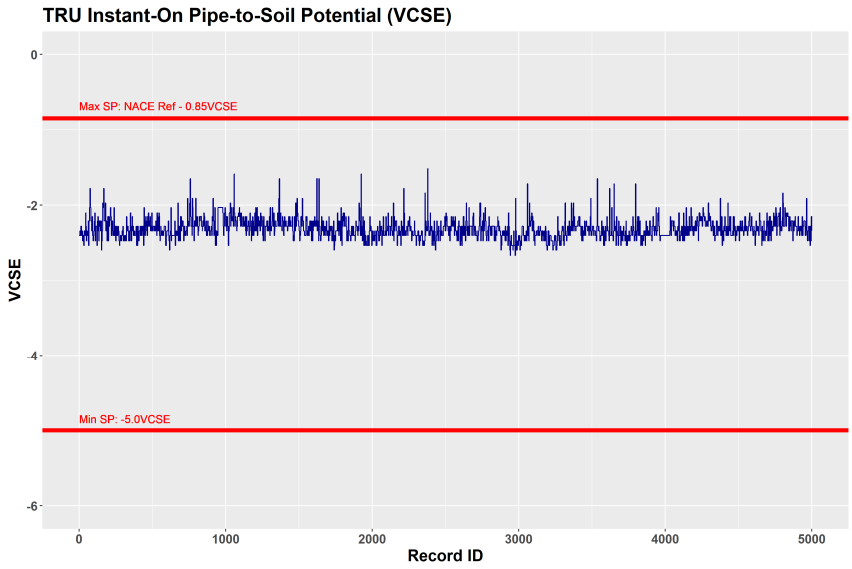

Figure 4-4 - TRU Instant-On CP pipe potential within OW Line Graph

4.3.1.4. Varying Pipe-to-Soil Potentials

This section investigates the presence of stray current and CP OP and UP exceptions at a TRU.

4.3.1.4.1. Pipe-to-Soil Potentials – Stray Current

The graph below illustrates the instant-on CP pipe potentials (V_{CSE}) from a TRU regulating the CP pipe potential with the presence of stray current (visible spikes).

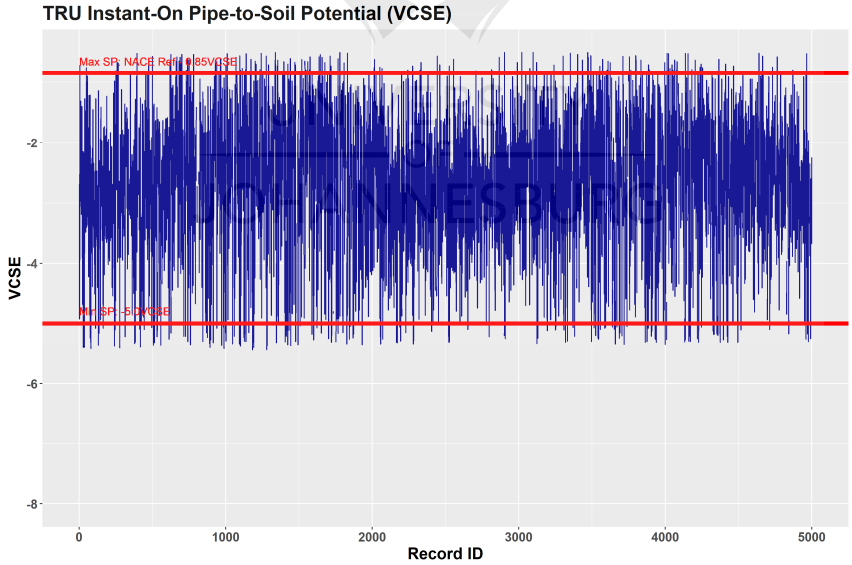

Figure 4-5 - TRU Instant-On Potential within OW & Stray Current Line Graph

The graph also indicates that the CP pipe potentials spikes through the maximum OW setpoint of -0.85V_{CSE} and the minimum OW of -5.0V_{CSE}. The resultant operating state is between the protection (P), over-protection (OP) and under-protection (UP) bands (when considering spikes as well). The magnitude of the %OP or %UP will depend on the statistical analysis for a specific period.

The literature review presented numerous possible sources for stray currents, such as DC transit systems, foreign pipelines or CP systems, telluric currents or AC-induced stray current. Stray current can pose a severe threat to a pipeline since the CP pipe potential can potentially spike in both the positive and negative directions.

The impact of stray current also relates to the condition of the coating, and a coating defect can result in a pipeline leak, where the significant stray current is present.

4.3.1.4.2. CP pipe potentials – Over Protection

The graph below illustrates the instant-on CP pipe potentials (V_{CSE}) from either a TRU supplying too much current and hence driving the CP pipe potential too electro-negative or significant stray current or interference is present. The resultant operating state is over-protection (OP).

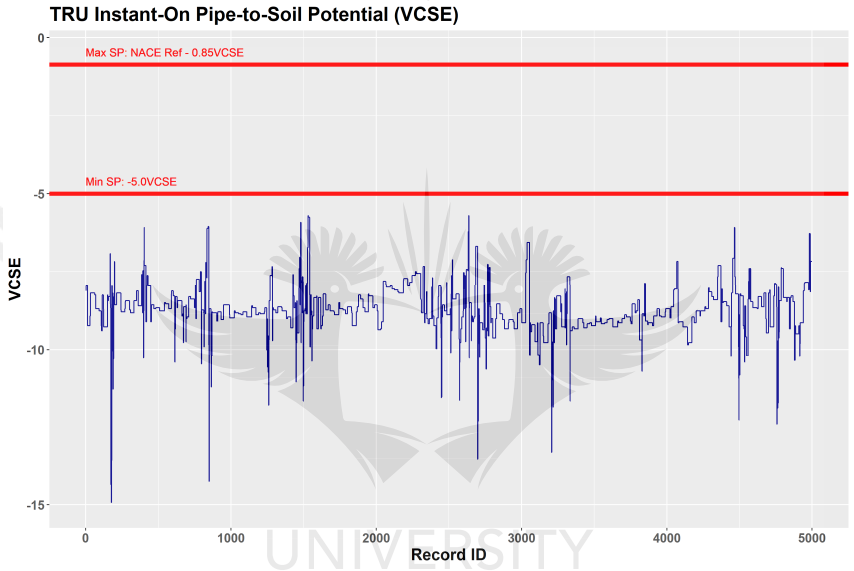

Figure 4-6 - TRU Instant-On Potential - Over-Protected Line Graph

As mentioned in the literature review, OP can lead to disbondment of the pipeline coating and this state is referred to as OP in this study. Stray current (visible spikes) is also present.

4.3.1.4.3. CP pipe potentials – Under Protection

Figure 4-7 illustrates the instant-on CP pipe potentials (V_{CSE}) for under-protection (UP).

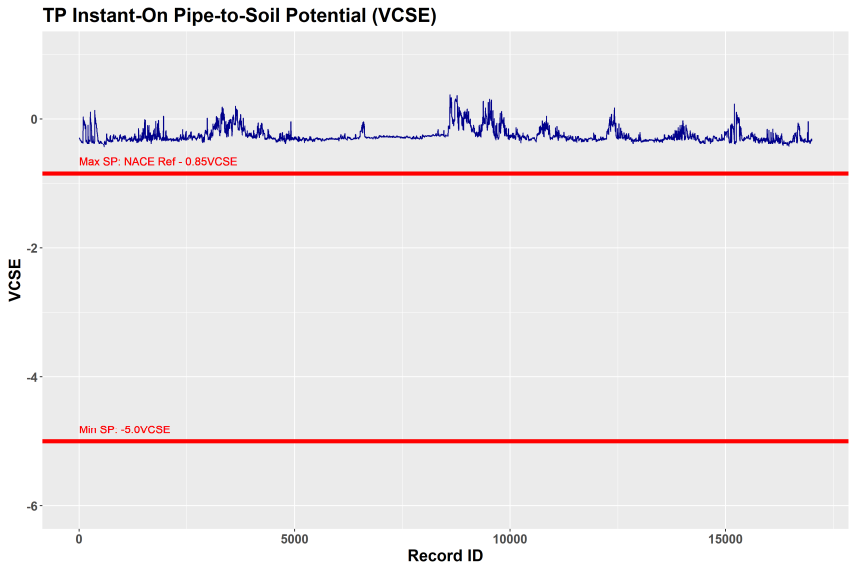

Figure 4-7 - TRU Instant-On Potential - Under-Protected Line Graph

Based on the NACE SP0169-2013 standard, UP is a severe threat to the pipeline over time, which can accelerate corrosion or result in forced corrosion.

4.3.1.5. Process Values (PV) Distribution

For the statistical distribution evaluation of the TRU PV values, a grid plot was created that indicates the density as well as the mean and median values:

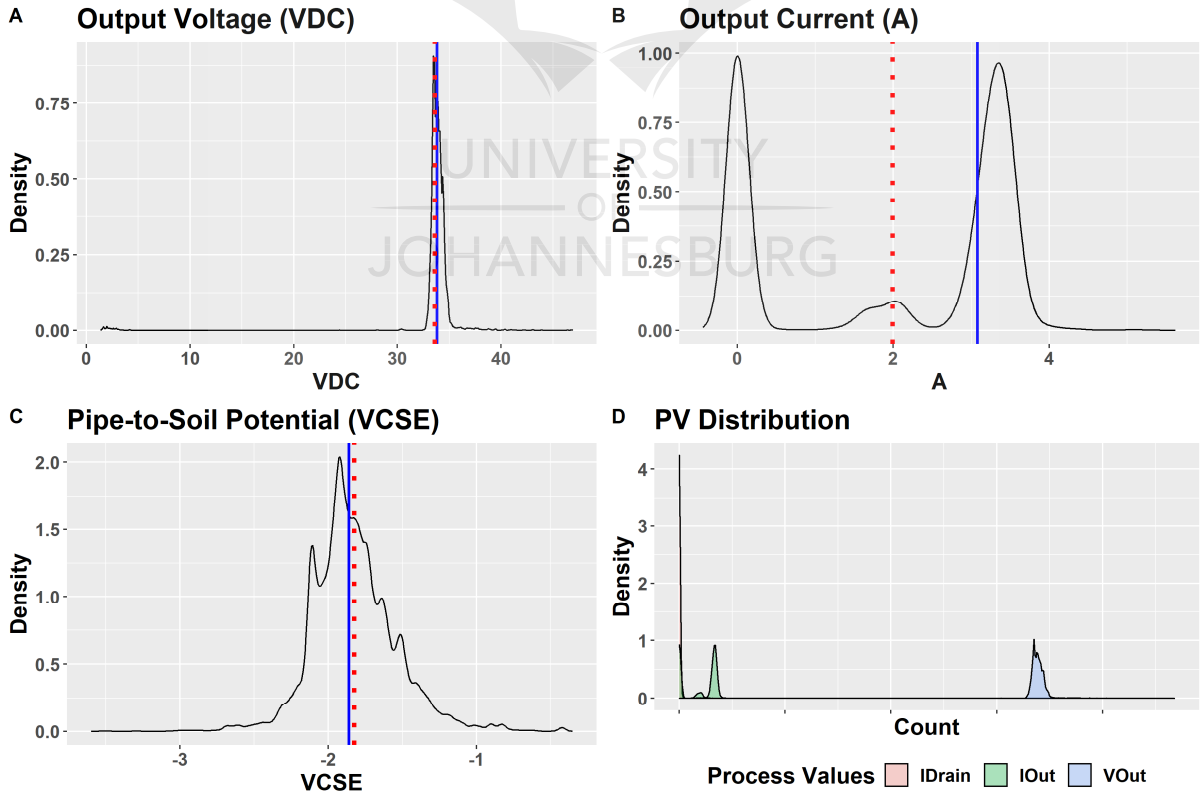

Figure 4-8 - TRU Process Value Distributions

CP pipe potential descriptive statistics can aid in fault-finding and inform maintenance activities and are discussed in the sections following.

4.3.1.6. PV Correlations

Correlation is a statistical method that seeks to determine linear relationships between continuous variables. The correlation coefficient is between -1 and +1, with 0 indicating no correlation between variables [81].

The corrplot library was used in R to determine if the PV's change as per the relevant theory (i.e. an increase in output current should result in an electro-negative shift in CP pipe potentials), and assuming that no stray current exists. Investigation of the PV value correlation provided the following results:

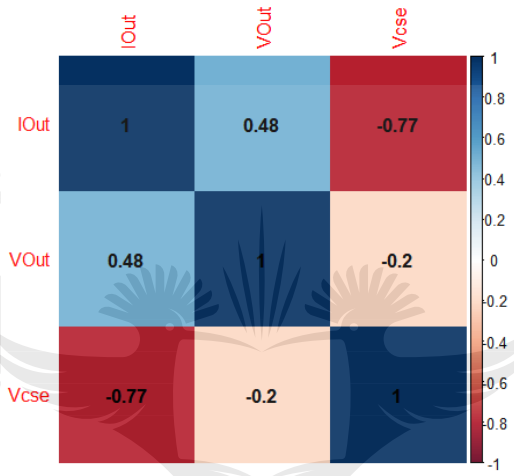

Figure 4-9 - TRU PV Correlation Plot - Regulating

From the correlation plot above, there is a strong negative correlation between the output current and the CP pipe potentials, whereas a moderate positive correlation exists between the output voltage and the output current. The table below illustrates the numeric correlation results:

TRU Correlation Results - Regulating		
Dataset	**Variables**	**Correlation**
Raw Data	Vcse vs IOut	-0.7834545
Raw Data	Vcse vs VOut	-0.1975518
Raw Data	VOut vs IOut	0.4687401

Table 4-2 - TRU Correlation Results - Regulating

When analysing a TRU where stray current exists, the correlation plot indicates weak to an extremely weak correlation between variables. The weak correlation is because stray current can dynamically shift CP pipe potentials positive or negative, and the rectifier might not have a change in output to counteract the change in CP pipe potential (due to malfunction, incorrect design or insufficient mitigation installed).

The table below illustrates the numeric correlation results.

TRU Correlation Results - Stray Current		
Dataset	**Variables**	**Correlation**
Raw Data	Vcse vs IOut	-0.006669
Raw Data	Vcse vs VOut	-0.3129182
Raw Data	VOut vs IOut	0.1531716

Table 4-3 - TRU Correlation Results – Stray Current

The correlation plot below indicates the variable correlation when evaluating a TRU with the presence of stray current:

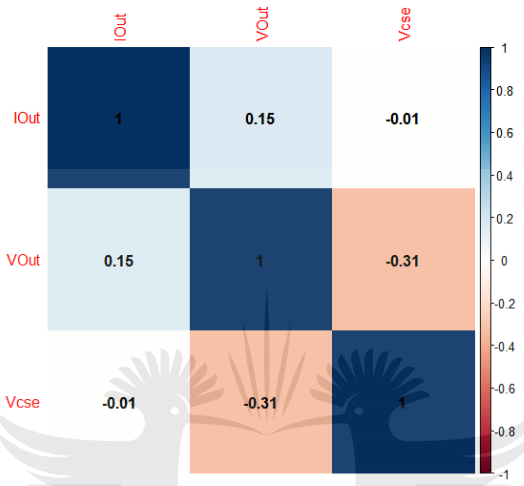

Figure 4-10 - TRU PV Correlation Plot - Stray Current

4.3.1.7. Time Series Decomposition of CP pipe potential

Time series analysis aims to define a mathematical model that describes the data set. Shumway and Stoffer suggest various models in R for time series analysis [106].

The decompose function of the R forecast package enables the programmer to retrieve the following components from time-series data [107]:

- Seasonal component – Represents the seasonal fluctuations based on time
- Figure – Mean seasonal effect
- Trend – Long-term increase or decrease in data
- Random – Random errors in data
- Type – Error type (multiplicative or additive)

To further analyse the trend, seasonality and error of the CP pipe potential data, a time-series object was created in R, and the object was decomposed into the different components.

The various components of the decomposed time-series are shown below and indicate a trend, seasonal and remainder component. The trend component can potentially be used to determine the CP pipe potential trend within a time window:

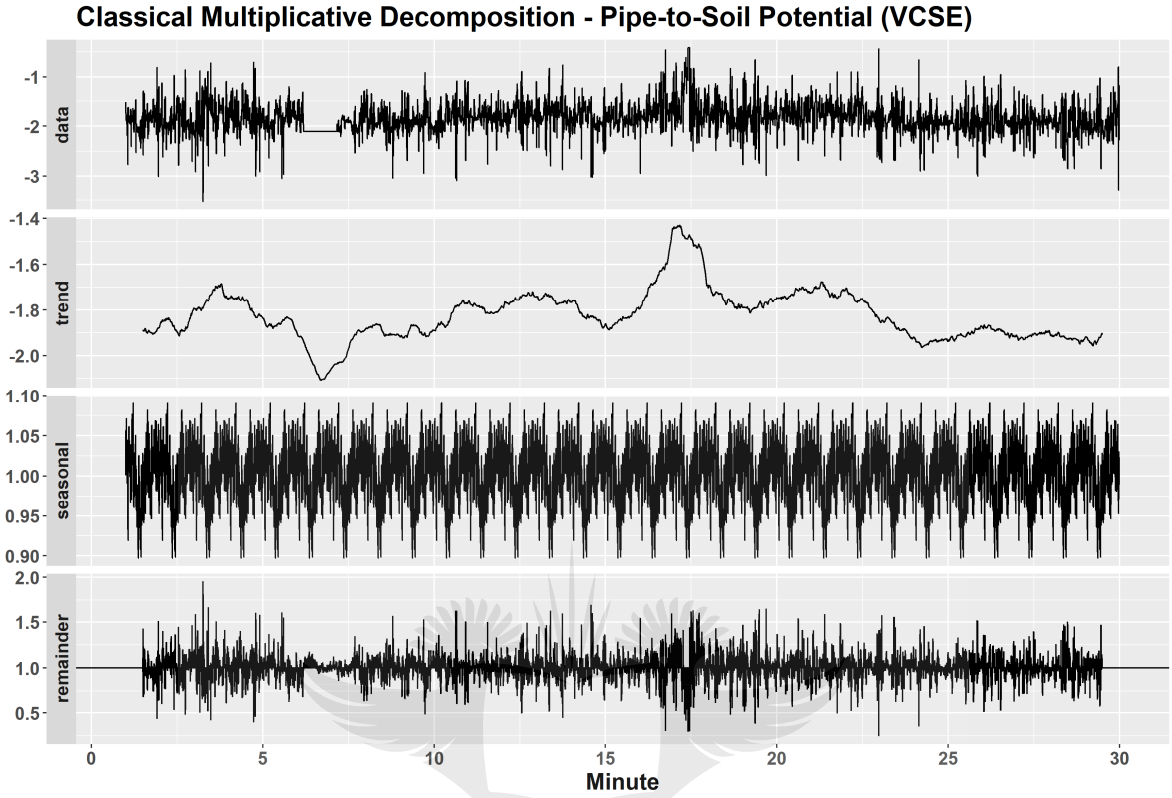

Figure 4-11 - Time-Series Decomposition of TRU CP pipe potential

A moving average (MA) is a statistical method that creates averages of subsets of the larger dataset. The MA is specified for a specific time and can also be considered for trend estimation. The graph below indicates the CP pipe potential with a 5-MA overlay:

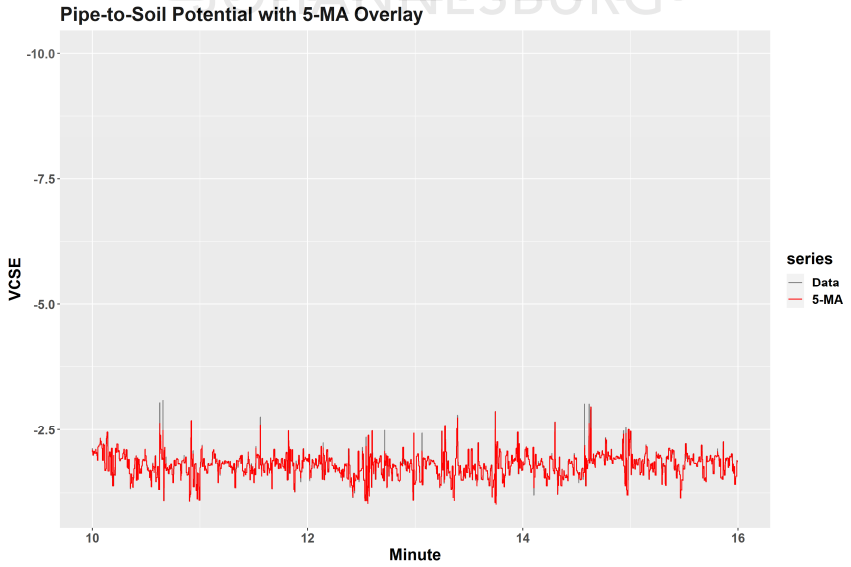

Figure 4-12 - TRU CP pipe potential vs 5-MA Line Graph

4.3.1.8. Section Summary

The TRU data analysis provided the following results:

- The CP pipe potentials can either operate within the OP, above, below or both. Data analysis should consider TRU operation within a period, rather than instantaneous monitoring of PV's (especially where stray current is present).
- The correlation between PV's change when stray current is present.
- The data distributions provide a visual clue as to the actual median operating values for each PV for the dataset time window.
- The trend component provides a trend for the CP pipe potential with noise eliminated.

The next section reviews the FDU data received.

4.3.2. FDU Data

4.3.2.1. Overview

Figure 4-13 represents a wiring diagram of a typical FDU, with the measurement points indicated in red:

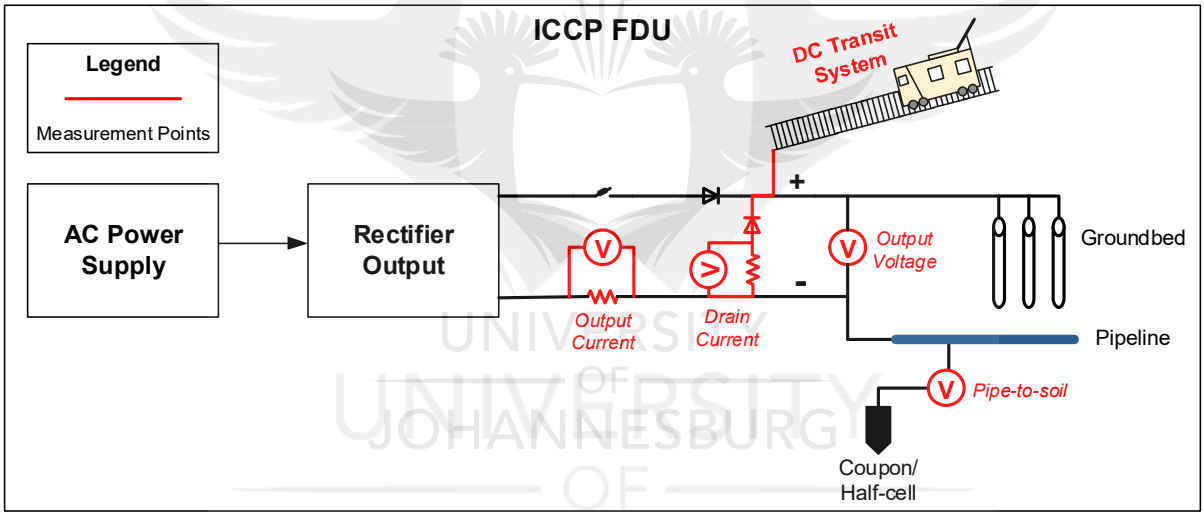

Figure 4-13 - FDU Circuit Diagram - Source: Adapted from [94]

Continuous data from an FDU received for this study consists of the following measurement points:

- Rectifier output voltage
- Rectifier output current
- Drainage current
- CP pipe potential (V_{CSE})

The operation of an FDU is similar to that of the TRU discussed in the previous sections. The only additional component is the diode (between the pipe and rail), used to drain stray current from the pipeline back to the DC transit system's rail.

When referring to the NACE SP0169-2013 standard for the criteria for CP protection, the measurement method needs to be recorded [11], [46]. For all datasets in this study,

no known IR drop exists, and the CP pipe potentials (V_{CSE}) are instant-on. A constant IR-drop factor can adjust the CP pipe potential to compensate for the IR-drop; however, no IR-drop compensation took place in this study (future work).

4.3.2.2. PV Evaluation for FDU's

Upon examining the raw data from the various FDU's, the steady-state PV's are theoretically supposed to vary within a defined control band (similar to the TRU and assuming no stray current exists). The drainage current can be constant, and spikes, when the train passes the FDU or the drainage current, is sporadic based on the transit system activity and or other stray current sources.

This section reviews the raw data received from the FDU's to determine variable correlation and investigate the FDU operation.

4.3.2.2.1. All FDU PV's

The graph below illustrates the rectifier output voltage and current, the drainage current, as well as the instant-on CP pipe potentials (V_{CSE}) from an FDU:

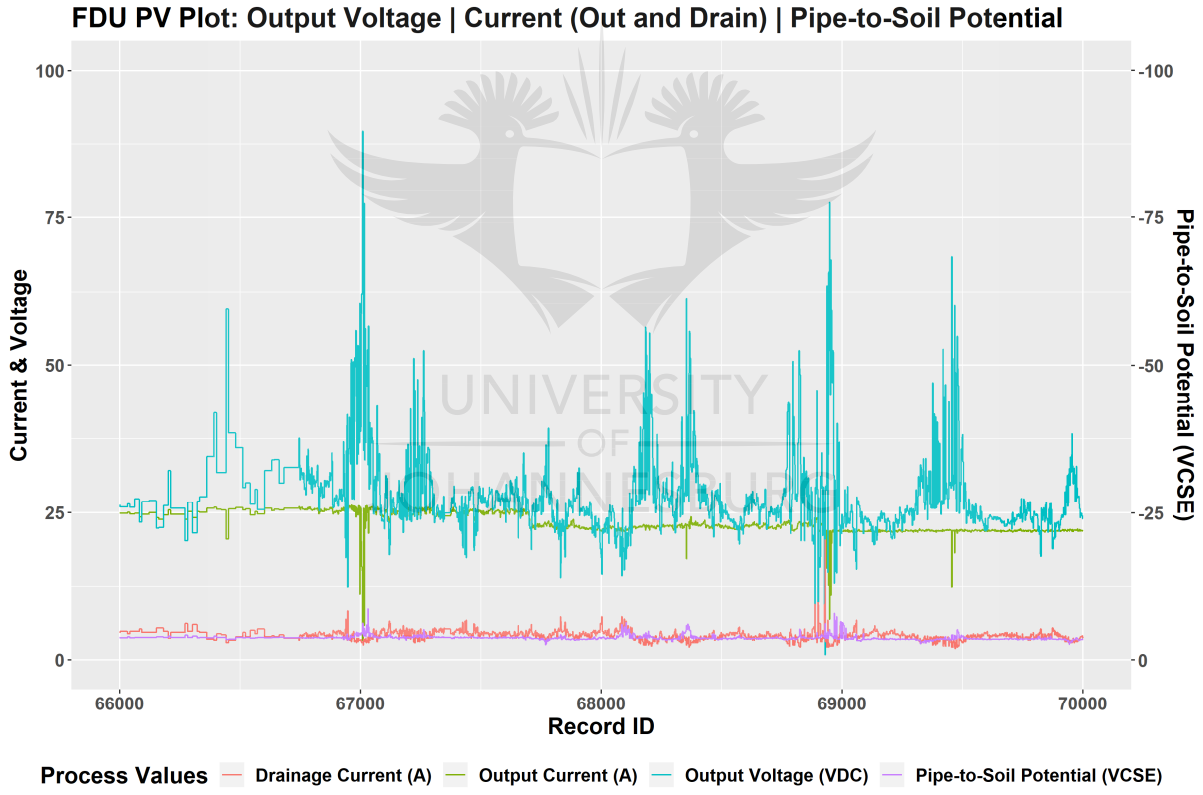

Figure 4-14 - FDU PV (All Measuring Points) Line Graph

From the graph above, the output voltage momentary spikes down, while the output current has a slight increase if the current drainage increase. After a short interval, a significant positive output voltage spike and negative output current spike is observed. These spikes are an attempt by the FDU to regulate the CP pipe potential with the presence of stray current.

4.3.2.2.2. CP pipe potential

Further investigation of only the CP pipe potential indicates that this specific FDU is not functioning 100% effectively due to the high magnitude negative CP pipe potential spikes when the current drainage increases.

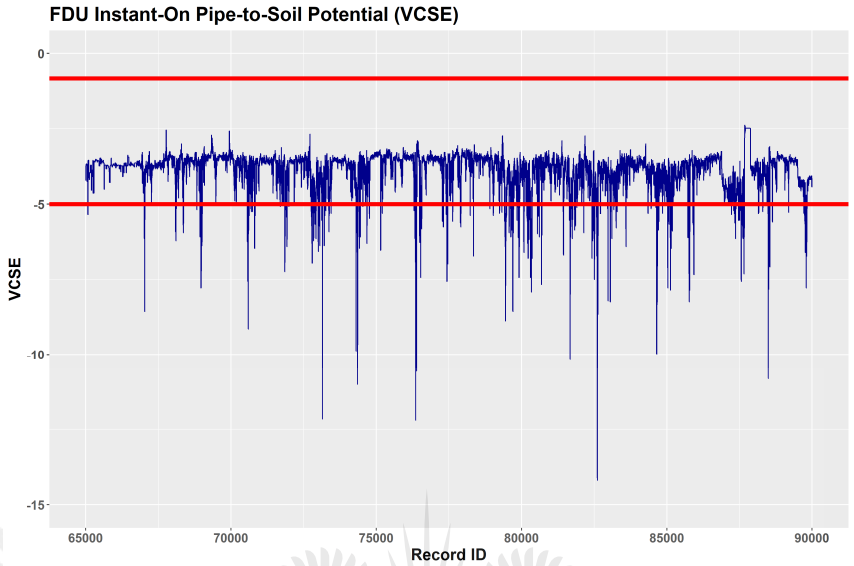

Figure 4-15 - FDU CP pipe potentials Line Graph

4.3.2.3. PV Distribution

For the statistical distribution evaluation of the FDU PV values, a grid plot was created that indicates the density as well as the mean and median values, as illustrated in Figure 4-16.

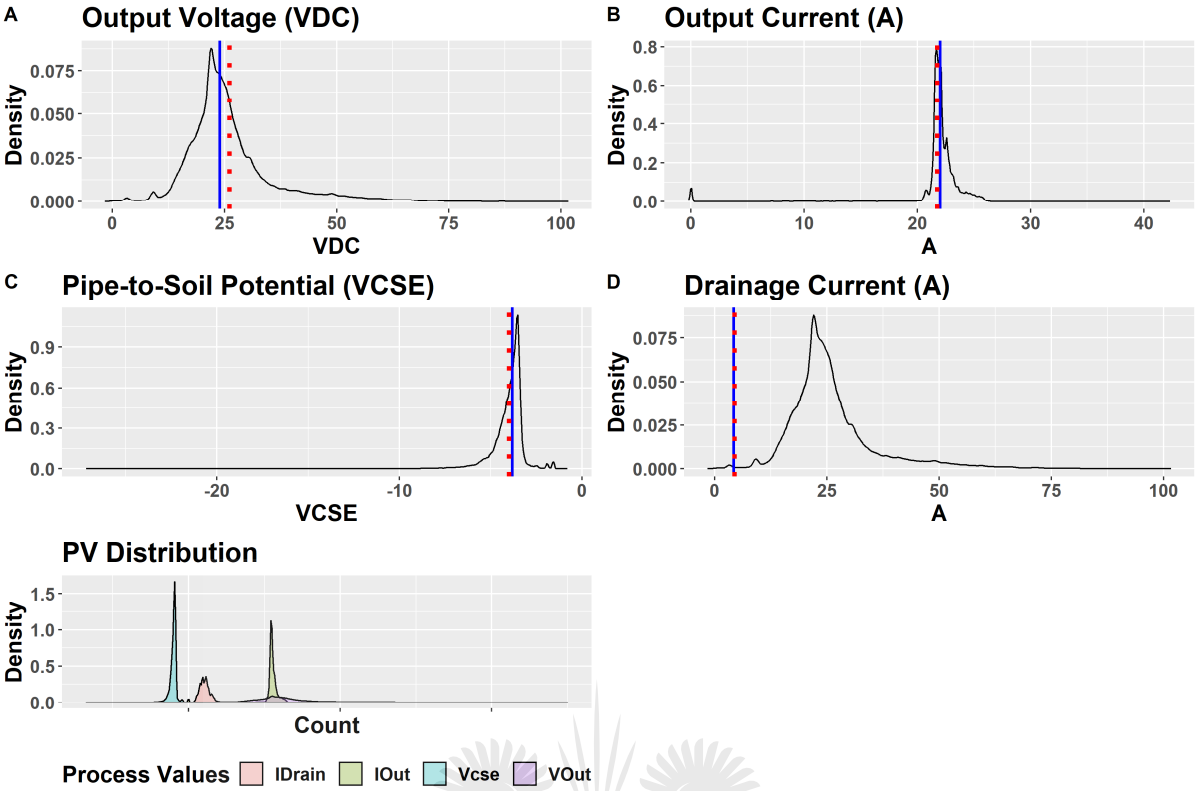

Figure 4-16 - FDU Process Value Distributions

Similar to the TRU, the data distributions provide visual clues of the FDU operation over time, rather than the instantaneous operation.

4.3.2.4. PV Correlation

Similar to the TRU analysis, the correlation function in R seeks to determine linear relationships between variables.

The correlation plot for this FDU is shown below:

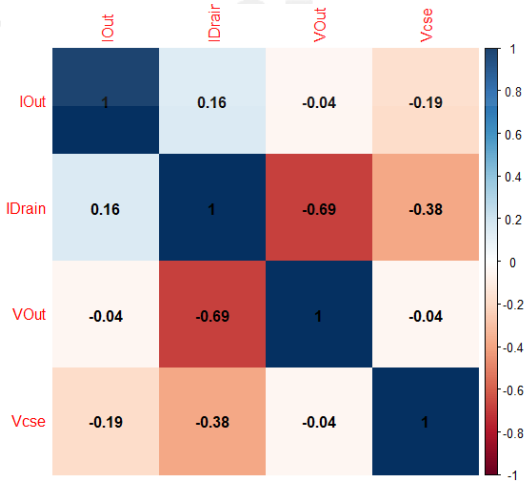

Figure 4-17 - FDU PV Correlation Plot

The correlation plot for the FDU supports the statement that the output voltage decrease if the current drainage increase (strong negative correlation), while there is a weak negative correlation between the CP pipe potential and drainage current.

The table below illustrates the numeric correlation results:

FDU Correlation Results		
Dataset	**Variables**	**Correlation**
Raw Data	Vcse vs IOut	-0.1894404
Raw Data	Vcse vs VOut	-0.0445657
Raw Data	VOut vs IOut	-0.0444984
Raw Data	Vcse vs IDrain	-0.3842015
Raw Data	VOut vs IDrain	-0.6862183
Raw Data	IOut vs IDrain	0.1574232

Table 4-4 - FDU Correlation Results

Härdle and Simar describe covariance as the relationship between random variables [108]. The covariance and correlation were evaluated to determine which is more applicable to the dataset for this study. To determine the covariance between variables, the eqs2lavaan package was used in R [109]. The resultant chart is a heatmap of the covariance and correlation matrix.

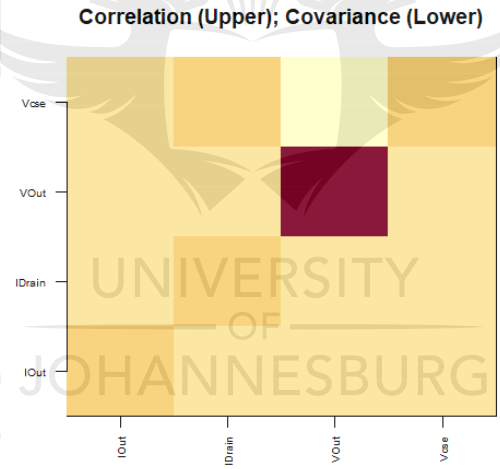

Figure 4-18 - Covariance Chart of FDU Variables

The results of the covariance chart align with the correlation matrix.

4.3.2.5. Time Series Decomposition of CP pipe potential

Similar to the time-series decomposition of the TRU CP pipe potential, the seasonal, trend and random error components for the FDU was analysed.

The components of the decomposed time-series are shown below and indicate the trend and seasonal components. The trend component can potentially be used for determining the CP pipe potential trend within a time window (as was the case for a TRU):

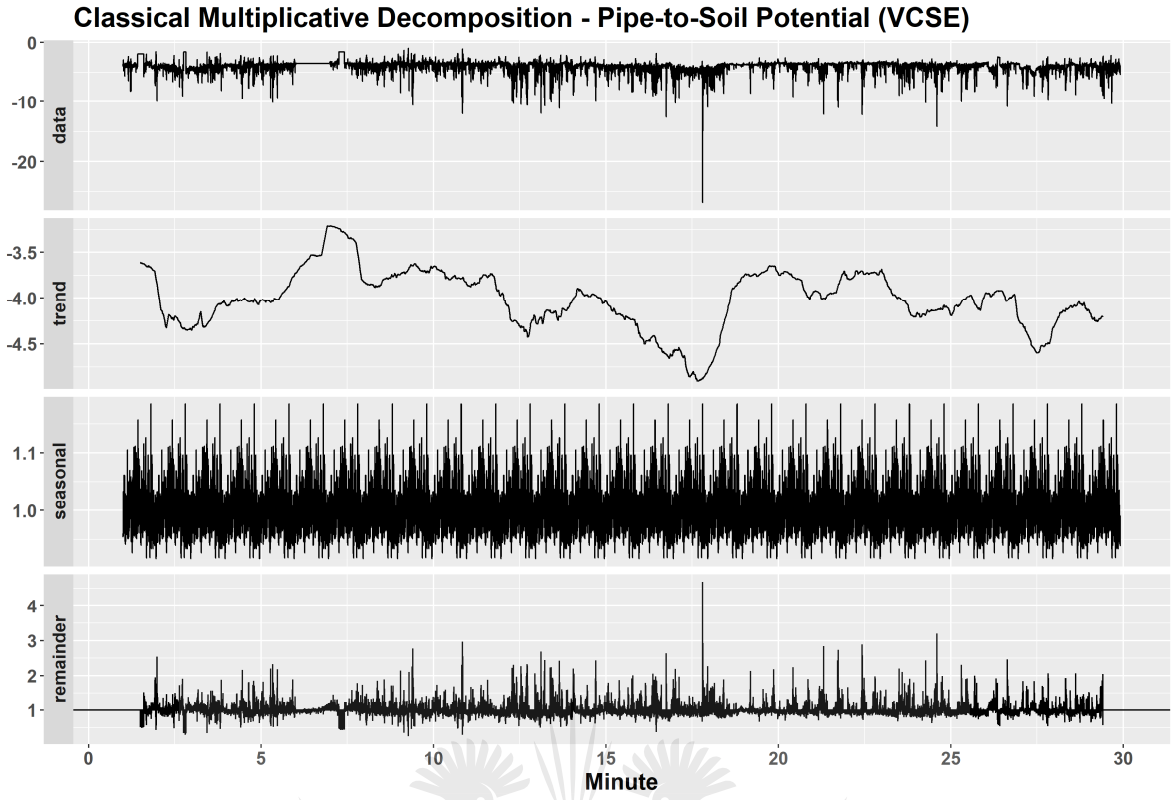

Figure 4-19 - Time-Series Decomposition of FDU CP pipe potential

Similar to using the trend component of the time series, a MA can also be considered for trend estimation. The graph below indicates the CP pipe potential with a 5-MA overlay:

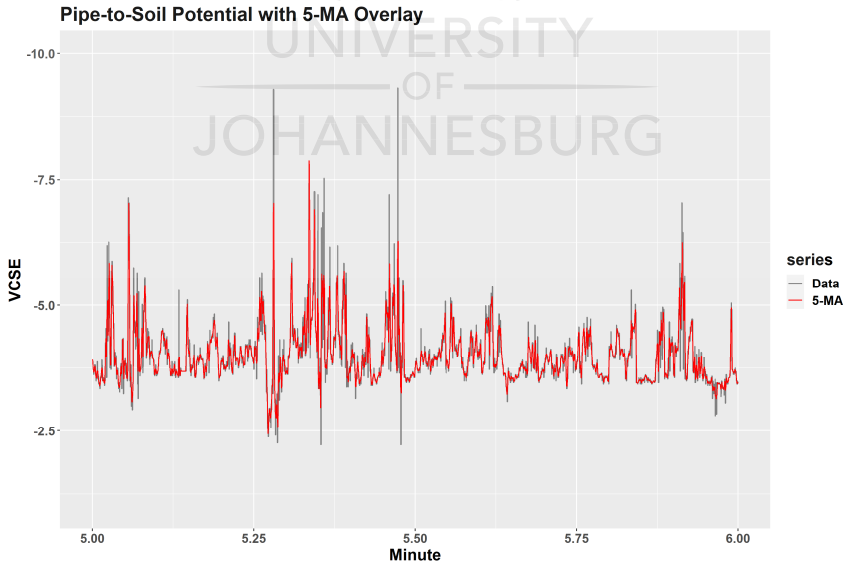

Figure 4-20 - FDU CP pipe potential vs 5-MA Line Graph

4.3.2.6. Section Summary

The FDU data analysis provided the following results:

- The CP pipe potentials can either operate within the OW, above, below or both. Data analysis should consider FDU operation over time, rather than instantaneous monitoring of PV's (especially where stray current is present).
- The correlation between PV's is different when compared to a TRU.
- The data distributions provide a visual clue as to the actual median operating values for each PV for the dataset time window.
- For CP pipe potential trend estimation, the time-series trend component or MA provides good results.

The next section reviews the CP pipe potentials for an FDU pipeline section.

4.3.3. FDU Pipeline Section

The sections following investigates the impact of discontinuous data on the analysis of CP pipe potentials, the CP health indicator and descriptive statistics.

4.3.3.1. Periodic Data

This sub-section evaluates an FDU pipeline section, which has continuous CP data for the FDU from the SCADA system, but only quarterly 24-hour recordings for the TP's following (recorded using a manual logger and on a set schedule). The pipeline visualization is as follows:

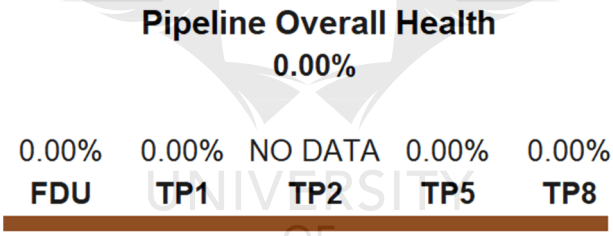

Figure 4-21 - FDU Pipeline Section

The TP number suffix indicates the kilometre distance from the FDU.

Determining the downstream effect of an FDU's operation, a line graph of the instant-on CP pipe potentials for the FDU and TP's provides a first glance visual overview:

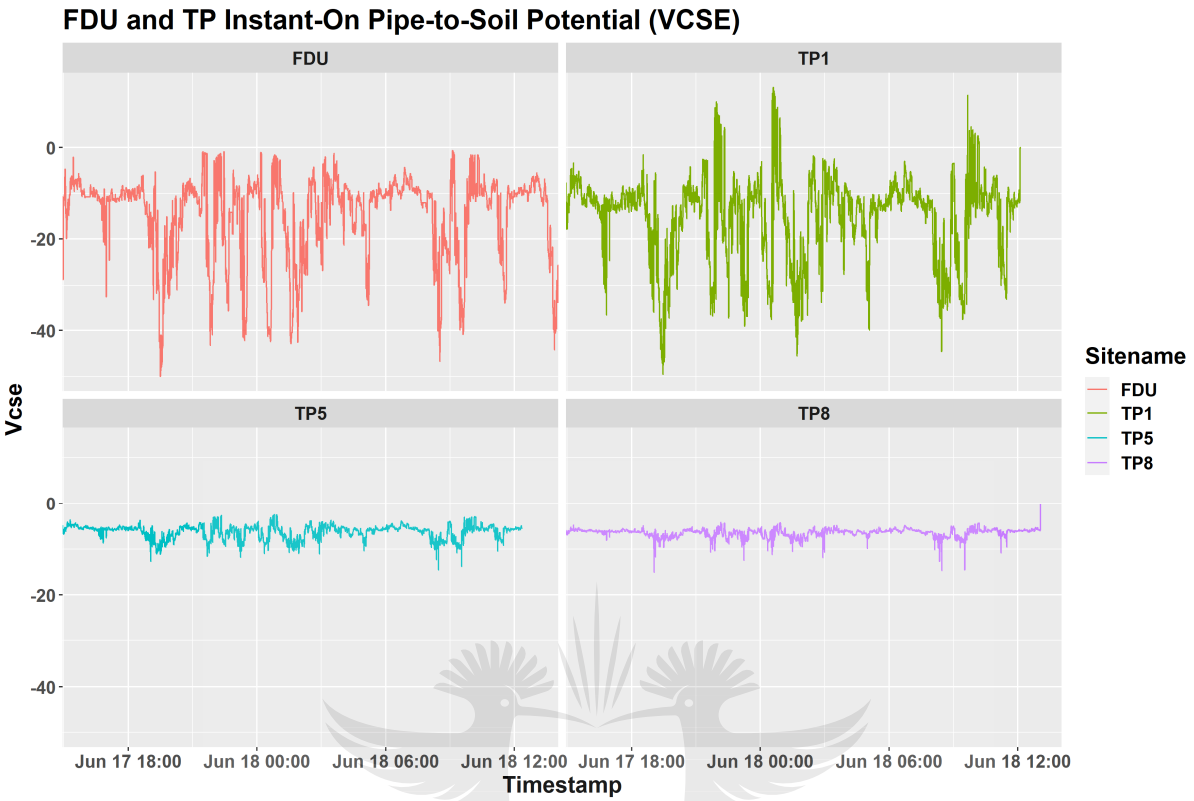

Figure 4-22 - FDU and TP CP pipe potential Comparison Graphs

From the sporadic CP pipe potential on the line graphs, it is evident that FDU either malfunctioned or stray current was present during the recording window. The sporadic CP pipe potential also affects the downstream TP's. The TP data follows the same trend; however, the magnitude of the spikes (up or down), reduces further along the pipeline (FDU spike -45V_{CSE}, while TP8 was at -18V_{CSE}).

Also visible from the line graphs above, is that TP data was only available for a short period (i.e. between 17 and 18 June). The periodic data measurement at TP's can potentially provide a false indication as to whether the CP is sufficient. The false indication can also occur if the rectifier was off, high stray current interference was present, or the rectifier was malfunctioning.

From the theory in the preceding sections regarding time-series analysis in R, the next step was to create a time-series object for each dataset and plot the trend for all four (through the decomposition of each time-series).

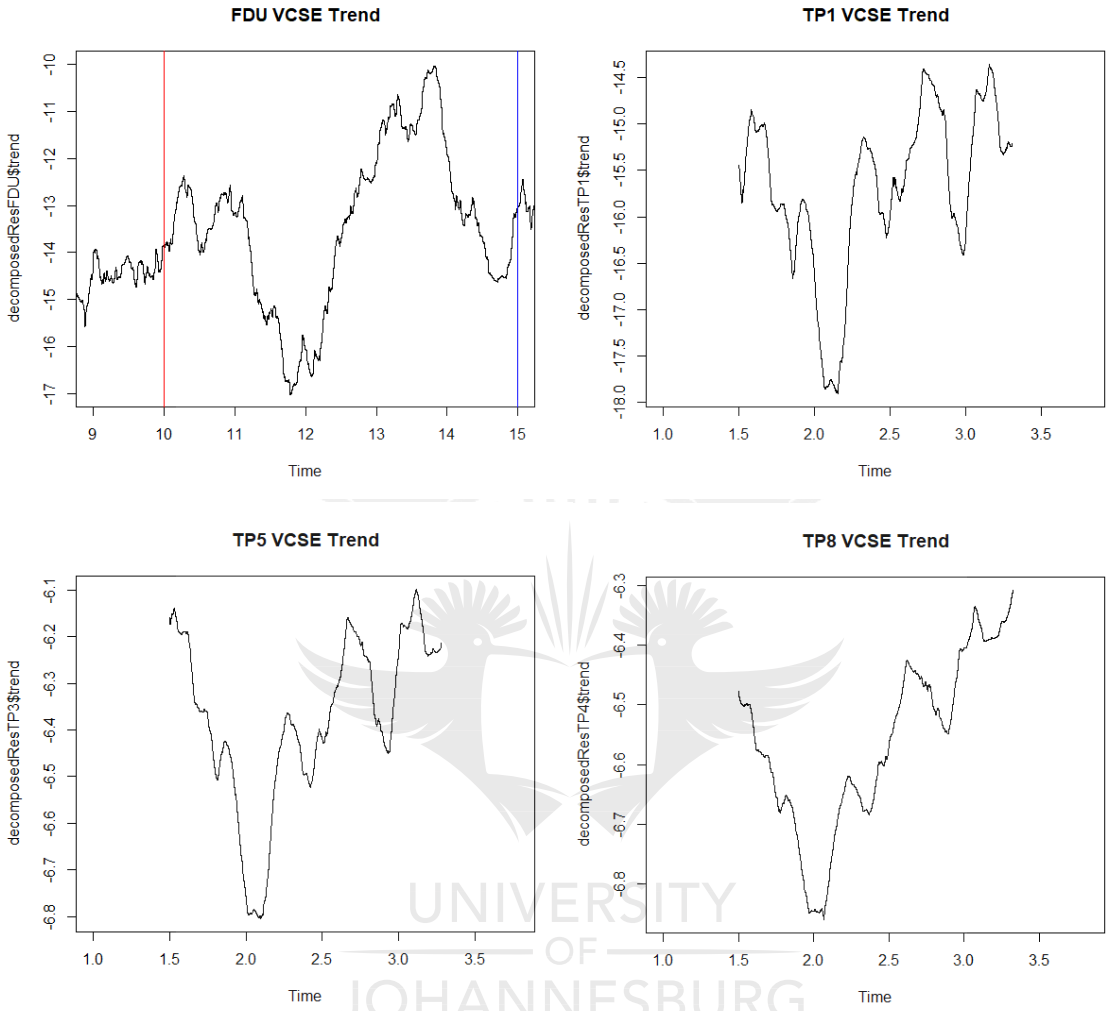

Figure 4-23 - FDU and TP CP pipe potential Trend Comparison Graphs

With the noise eliminated from the time-series object, the trend component for each TP indicates the potential shift at each TP and the relative stray current decay along the pipeline.

For evaluation of the CP health indicator described in chapter three and Appendix B, two scenarios were considered for the health indicator calculation. Firstly, the health indicator was calculated based on the assumption that there was no TP data available (only FDU data available) and secondly; the periodic data for three TP's was available (TP1, TP2 and TP8).

The table below provides the CP health indicator data for three scenarios evaluated:

CP Health Indicator Results						
FDU	**TP1**	**TP2**	**TP5**	**TP8**	**Overall Pipe CP Health**	**Test**
62.570%	74.780%	77.204%	85.898%	82.181%	76.358%	With Data
62.570%	86.136%	77.204%	79.974%	75.353%	76.009%	Approximation
62.570%	74.671%	77.204%	82.465%	85.753%	76.365%	Approximation - Inverse

Table 4-5 - CP Health Indicator Results (Periodic Data)

The first row displays the results where the data was available for the calculation of the CP health indicator and indicate that the stray current effects at the FDU decay on the TP's down the line. The second and third-row indicates the results of the CP health indicator where no TP data was available, and the CP health indicator was calculated using formula 3.4. TP2 was calculated based on the approximation approach below due to the unavailability of data.

From the two risk level formulas(3.1 and 3.2) presented in chapter three, the inverse approximation is most applicable here, as the stray current effects reduce along the pipeline. This approximation also includes a % error at TP3 and TP4, which is accurate to 3% for this example.

Since most of the importance of CP data analysis is placed on CP pipe potentials, the descriptive statistics per rectifier and TP is suggested as described in chapter three (descriptive statistics such as min/max/average and others). These results are tabled below:

CP Descriptive statistics										
Unit	**Min V_{CSE}**	**Max V_{CSE}**	**Avg V_{CSE}**	**Range V_{CSE}**	**% P**	**% UP**	**% OP**	**Time P (mins)**	**Time OP (mins)**	**Time UP (mins)**
FDU	-50.00	0.48	-14.82	50.48	4.80	0.05	95.14	2006.50	39731.41	22.09
TP1	-49.56	13.25	-15.49	62.81	3.11	2.60	94.29	41.83	1269.34	35.00
TP5	-14.52	0.02	-6.27	14.54	12.38	0.10	87.52	165.15	1168.02	1.33
TP8	-15.10	0.01	-6.47	15.11	1.54	0.02	98.44	20.83	1334.34	0.33

Table 4-6 - CP Descriptive statistics (Periodic Data)

From the descriptive summary statistical results above, the following was observed:

- The average CP pipe potential provides a baseline CP pipe potential for the dataset period and can potentially be used to report performance.
- The range statistic describes the significance between the minimum and maximum CP pipe potential and can indicate the presence of stray current or interference.
- The percentage and time statistics describe the CP pipe potential conformance to the OW.

The above statistical performance for the dataset period can aid in determining the maintenance activity to remedy the CP-related problem.

Below is a typical graphical representation of the pipeline's CP health and descriptive statistics calculated:

Pipeline Overall Health

76.530%

% OP	95.14%	94.29%		87.52%	98.44%
% UP	0.05%	2.60%		0.10%	0.02%
% P	4.80%	3.11%		12.38%	1.54%
Health	62.570%	74.780%	77.204%	85.898%	82.181%
	FDU	**TP1**	**TP2**	**TP5**	**TP8**

Figure 4-24 - Pipeline Overall Health and Descriptive statistics

The time and percentage statistics and pipeline health in the above example are not correlated, because the time and percentage statistics only classify the CP pipe potential as either operating in three bands (namely, P, OP and UP), while the CP health indicator, considers four risk levels, based on the defined CP pipe potential bands for P, OP and UP.

To improve the relationship between the descriptive statistics and the CP health calculation, two possible options exist, namely, reducing the potential-band (PB) per risk level or combining the CP health indicator with the time and percentage statistics.

The results below show the pipeline health with reduced risk levels per OW:

Pipeline Overall Health

70.630%

% OP	95.14%	94.29%		87.52%	98.44%
% UP	0.05%	2.60%		0.10%	0.02%
% P	4.80%	3.11%		12.38%	1.54%
Health	56.190%	70.330%	73.300%	79.460%	72.310%
	FDU	**TP1**	**TP2**	**TP5**	**TP8**

Figure 4-25 - Pipeline Overall Health and Descriptive Statistics (Adjusted OW)

The health of the pipeline is now lower with the reduced risk levels per OW, although not highly correlated with the time and percentage statistics. To present the pipeline health as a combination of the time and percentage statistics, the candidate suggests using both metrics for maintenance decision-making and CP health evaluation.

The addition of conditional colour formatting improves the usability of the pipeline section below:

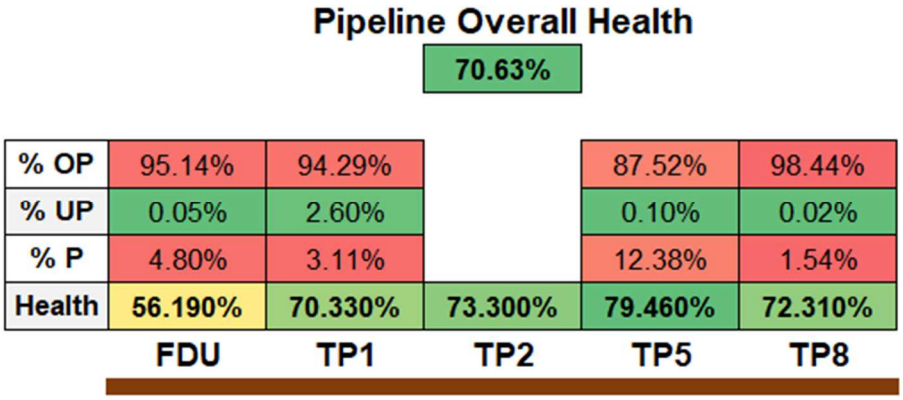

Figure 4-26 - Pipeline Overall Health and Descriptive statistics (Colour Formatting)

The next section explores the FDU operation, where continuous remote data is available.

4.3.3.2. Continuous Data

This sub-section evaluates an FDU pipeline section, which has continuous CP data from data loggers for the TP's downstream of the FDU. The FDU is also continuously monitored using a SCADA system. For this evaluation, data were analysed for the last 30 days.

Similar to the preceding section, to evaluate of the CP health indicator described in chapter three and Appendix B, two scenarios were considered for the health indicator calculation. Firstly, the health indicator was calculated based on the assumption that there was no TP data available (only FDU data available) and secondly; continuous data for four TP's was available.

CP Health Indicator Results						
FDU	**TP1**	**TP2**	**TP3**	**TP4**	**Overall Pipe CP Health**	**Test**
55.940%	99.990%	69.040%	99.870%	74.280%	79.828%	With Data
55.940%	84.282%	83.070%	81.864%	70.980%	75.229%	Approximation
55.940%	69.030%	71.250%	76.323%	70.980%	71.156%	Approximation - Inverse

Table 4-7 - CP Health Indicator Results (Continuous Data)

Based on the results above, the CP pipe potentials of TP2 and TP4 seem to fluctuate with the FDU output (with decay). TP1 and TP3 present high overall health (99%), which does not correlate with the FDU output. The candidate assumed that TP1 and TP3 had an error recording the correct potentials or the results are inconclusive of error.

The descriptive statistics for continuous data is displayed below:

CP Descriptive Statistics										
Unit	**Min V_{CSE}**	**Max V_{CSE}**	**Avg V_{CSE}**	**Range V_{CSE}**	**% P**	**% UP**	**% OP**	**Time P (mins)**	**Time OP (mins)**	**Time UP (mins)**
FDU	-50.00	0.48	-14.82	50.48	4.80	0.05	95.14	2006.50	39731.41	22.09
FDU	-50.00	-0.10	-15.23	49.90	7.74	0.07	92.18	3233.97	38495.85	30.18
TP1	-4.50	-0.53	-2.20	3.98	0.02	0.00	99.98	14.20	91535.80	0.00
TP2	-43.87	2.14	-12.84	46.01	3.24	0.96	95.79	2970.44	87696.46	883.10
TP3	-5.91	0.00	-2.76	5.91	99.25	0.64	0.10	90863.57	96.00	590.43
TP4	-22.53	1.06	-7.46	23.59	13.67	0.14	86.20	12513.82	78912.04	124.14

Table 4-8 - CP Descriptive Statistics (Continuous Data)

The results from the table above indicate, the performance of the FDU and the TP's over a more extended period, which includes more data points and is more representative of the CP pipe potentials over time and the effectiveness of the CP system in comparison to the use of periodic data.

The candidate suggests that systematic and periodic analysis of descriptive CP statistics can improve the maintenance response as well as describe the performance over extended periods.

4.3.3.3. Summary

The FDU pipeline section data analysis provided the following results:

- The CP health indicator indicates the CP health based on the criteria defined section three (risk, unit type, location and potential bands), whereas the descriptive statistics describes the unit performance based on the three operating bands of P, OP and UP.
- Continuous data provides an improved view of the pipeline CP system health, whereas periodic data, only takes a snapshot of the current conditions.
- The CP health indicator can potentially be used to determine the health of downstream TP's where no continuous data is available for TP's. However, the OP, UP and P factors need to be calibrated per pipeline section to improve reliability.
- The descriptive statistics can be used for long-term comparative analysis and inform the maintenance required.

The next section explores long-term time-series analysis applicable to the scope of this study.

4.3.4. Long-Term Time Series Analysis

Evaluating the CP pipe potentials over a more extended period (more than 12 months) provides useful information about the trend of the CP pipe potentials, which can, in turn, inform the required long-term or seasonal maintenance approach. For the graphs in this section, continuous data was collected from a data logger installed at a TP.

The raw data was read in R, and a time series object was created for plotting the decomposed time-series objects. Four frequencies were specified for the time-series objects, namely, hourly, daily, weekly, monthly and quarterly. The results are discussed below.

4.3.4.1. Hourly Trend

The time series plot objects in R allows one to plot the CP pipe potential for a specific seasonality period or frequency (with all the noise removed). The left plot's frequency was set for one hour. The plot is, however, not very user-friendly and filtering of the Time axis is suggested for analysis (plot on the right). Hourly analysis can aid in performing CP investigations.

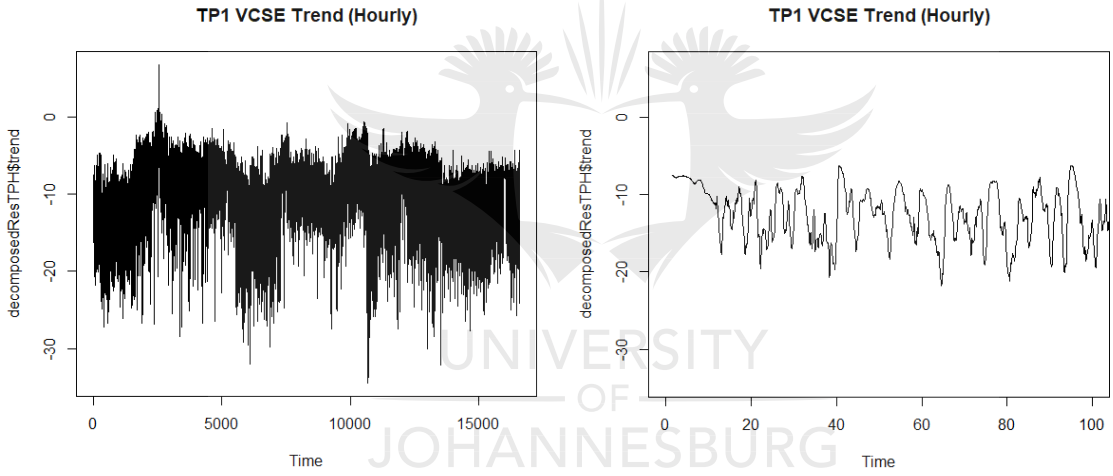

Figure 4-27 - Long-term CP pipe potentials - Hourly Trend Line Graphs

4.3.4.2. Daily Trend

The daily trend plots the daily CP pipe potential value with all the noise removed. For preventive maintenance, forecasting techniques can be used to estimate the value of the CP pipe potential for the next day, and personnel can potentially be scheduled accordingly.

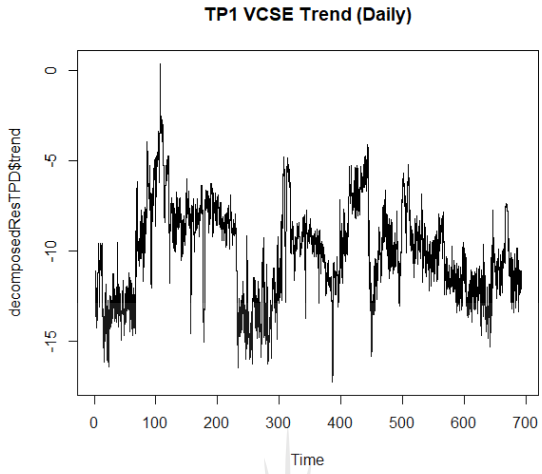

Figure 4-28 - Long-term CP pipe potentials - Daily Trend Line Graph

A basic forecast plot is shown below that considers three forecasts methods in R, namely, mean, naïve and season naïve:

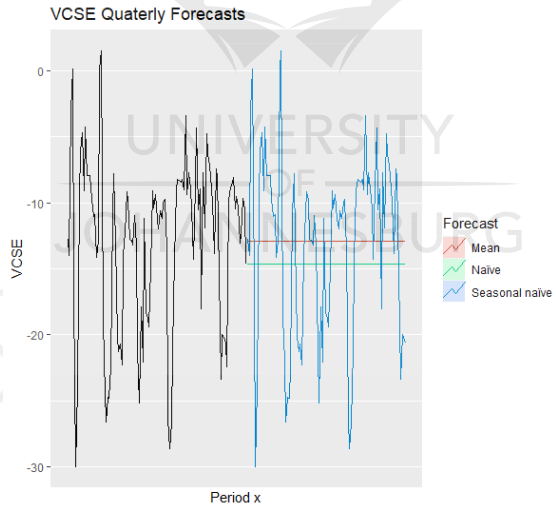

Figure 4-29 - Long-term CP pipe potentials with Forecast Line Graph

4.3.4.3. Weekly Trend

If the granularity of the daily trend is too high for maintenance operations, the weekly trend can be used for maintenance scheduling.

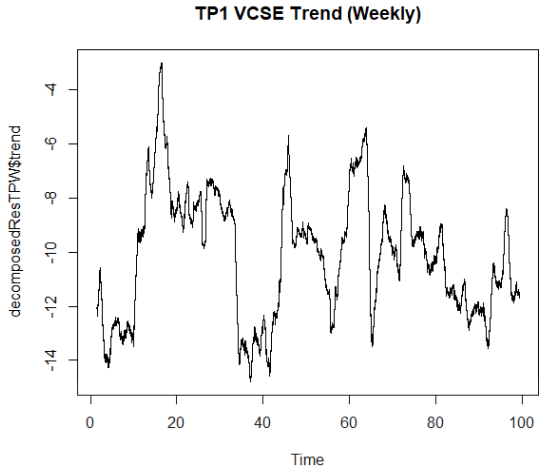

Figure 4-30 - Long-term CP pipe potentials - Weekly Trend Line Graph

4.3.4.4. Quarterly Trend

The quarterly trend can inform the effect of seasonal weather changes on the CP pipe potentials and aid as a tool to adjust rectifiers before seasonal weather changes.

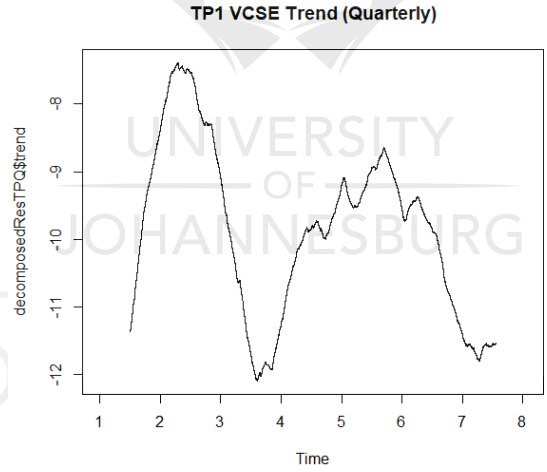

Figure 4-31 - Long-term CP pipe potentials - Quarterly Trend Line Graph

4.3.4.5. Summary

This section reviewed the time-series decomposition for specific intervals, intending to provide possible use-cases. If the data is available for extended periods, the continuous analysis can inform the maintenance required.

4.3.5. Conclusion

This chapter focussed on performing an exploratory data analysis, mainly focussing on CP pipe potentials for a TRU and FDU, as well as an FDU pipeline section. The

TRU and FDU section evaluated the CP pipe potentials relative to operation within the defined OW's and its correlation with regards to the rectifier's output voltage, output current and drainage current (FDU only). It was evident that the presence of stray current affects the correlation of rectifier variables monitored. This finding translates to a conclusion that the absence of mathematical modelling of rectifier operation operating with stray current, can pose a significant challenge for predictive approaches.

The FDU pipeline section evaluated the CP health indicator and the descriptive statistics applicable to this study. The CP health indicator provided overall pipeline health accurate to approximately 4-5% (with FDU data only). It was however evident that the risk-level factors need to be adjusted per rectifier (which in practice can be related to dynamic changes of external conditions). Furthermore, the combination of the descriptive statistics and the CP health indicator is suggested to evaluate the health of the CP system.

The FDU pipeline section was also evaluated with periodic and continuous data, and the advantage of continuous data is evident when an analysis is performed for more extended periods. The periodic data also poses a problem as it just describes the CP pipe potentials for the specified recording window and might be error-prone due to the rectifier condition at the time of recording.

Lastly, the analysis of the trend component of a time-series object provided mechanisms that can be used for CP system analysis and possibly inform the required maintenance activity.

The next chapter discusses the results of the ML model and survival time analysis.

5. CHAPTER 5: PREDICTIVE MODELLING EVALUATION AND RESULTS

5.1. Introduction

Building on the exploratory data analysis findings, the focus of this chapter is to evaluate the predictive modelling results, which includes the TRU/FDU state prediction, predicting the effect on downstream TP's, evaluating the time-to-state analysis and lastly evaluating the maintenance suggestion capability of this study.

This chapter consists of the following sections:

i. Evaluate the performance of four ML models for predicting the pipe's potential state of either a TRU or FDU
ii. Explore the predictive results of the downstream effect on TP's
iii. Evaluate the time-to-state analysis for a defined state prediction
iv. Evaluate the suggested maintenance capability of this study

Similar to chapter four, predictive modelling was performed using the R Studio IDE.

5.1.1. Key Terminology

The applicable terminology for this chapter includes [81]:

- Predictors – Input variables for prediction equation
- Outcome – Result of the prediction equation
- Categorical data – Data with discrete values
- Sample – single or subset of data

5.2. CP pipe potential Prediction

As discussed in chapter four, the CP pipe potential of a TRU, FDU or TP is the most important measurement that is taken for evaluating the effectiveness of a CP system (guided by the protection criteria of the NACE SP0169-2013 standard). This section discusses the ML modelling approach and evaluates the performance for predicting the CP pipe potential at either a TRU, FDU or TP.

5.2.1. ML Performance Evaluation

Chapter three considered the metric that will be used for evaluating the prediction accuracy of an ML model in this study. The RMSE or MAE are standard metrics available in R for this evaluation and will be referenced throughout this chapter.

5.2.2. Evaluation of Predicted CP pipe potentials

Decomposition of the CP pipe potential time-series objects in chapter three presented the possibility to forecast future values based purely on the historical values of the said time-series. This section, however, evaluates the prediction of the CP pipe potential for a TRU or FDU, based on predictor variables such as the rectifier output voltage, output current and drainage current (FDU only).

5.2.2.1. Steady-State Operation

When a system is in steady-state, the variables describing the system have minor or no changes over time [110]. The modelling and prediction results of CP pipe potentials of a TRU operating at steady-state, i.e. regulating the CP pipe potential within the OW, are discussed in this section. The raw data received from the SCADA system has a sampling interval of 30 seconds and a period of 30 days.

5.2.2.1.1. Modelling Approach

The ML modelling approach consisted of the following steps:

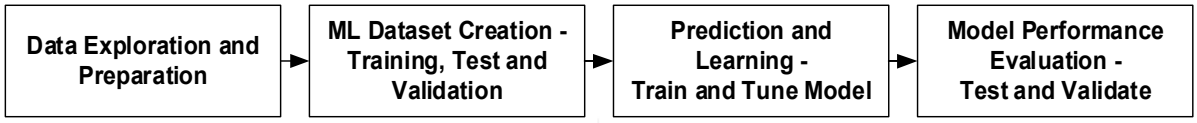

Figure 5-1 - ML Modelling Approach - TRU Steady-State

The dataset exploration and preparation included the removal of erroneous rows, column formatting and an evaluation of the skewness, outliers and centring and scaling of variables, as suggested by Kuhn and Johnson [81].

5.2.2.1.1.1. Data Exploration and Preparation

The skewness results for the dataset indicates the asymmetry, whereby the CP pipe potential is skewed to the left and the output voltage and current are skewed to the right:

TRU Skewness Results		
Dataset	**Variables**	**Skewness**
Original	Vptg	-0.4277091
Original	VOut	0.418736
Original	IOut	0.424159

Table 5-1 - TRU Skewness Results

For evaluation of outliers, a boxplot was created for the CP pipe potential, rectifier output voltage and current. The boxplot results are summarized as follows:

- The output current fluctuates between 0A and 10A (inter-quantile range), with a median value at around 1A. An upper whisker is present at around 12A.
- The output voltage fluctuates between 3V and 25V (inter-quantile range), with a median value of 4V. A lower and upper whisker is present at approximately 2V and 28V.
- The CP pipe potential is not visible, and a separate boxplot is required for this variable only.

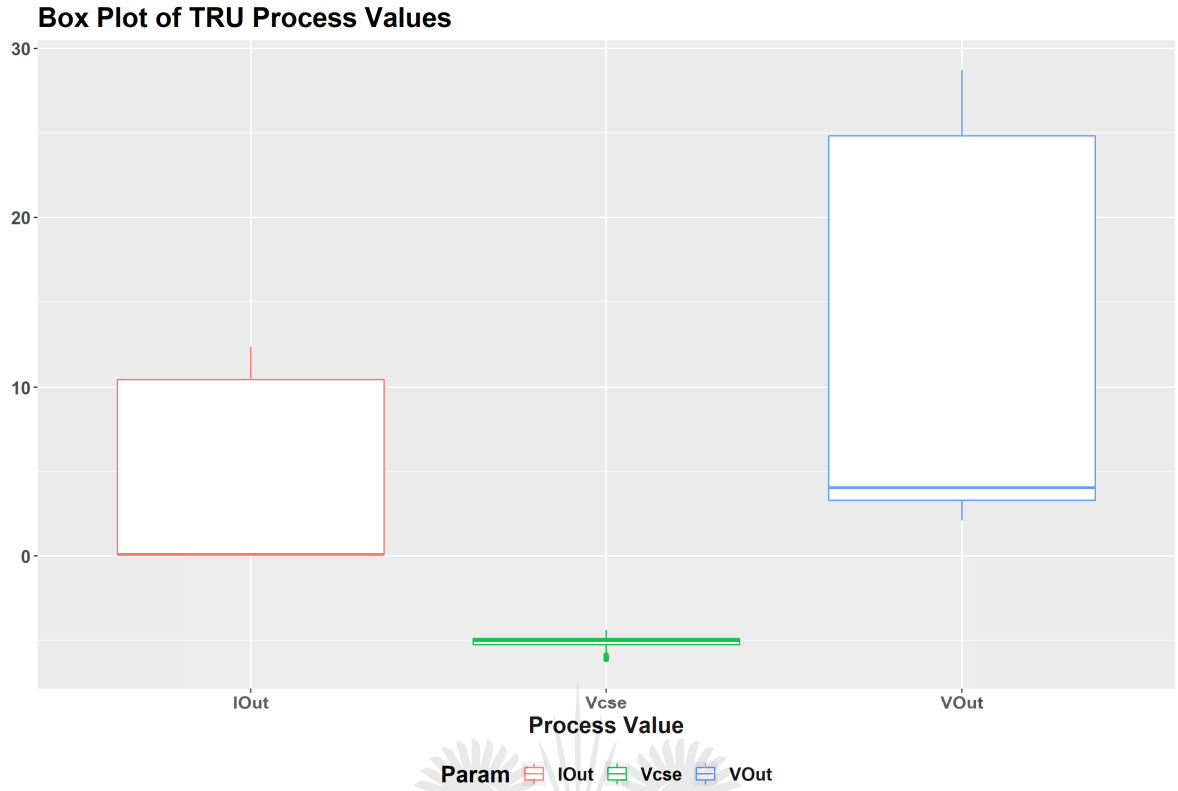

Figure 5-2 - Boxplot of TRU PV's

Based on the boxplot above, only the CP pipe potential have outliers that can affect the data analysis. The outliers are visible on the boxplot for the CP pipe potential only:

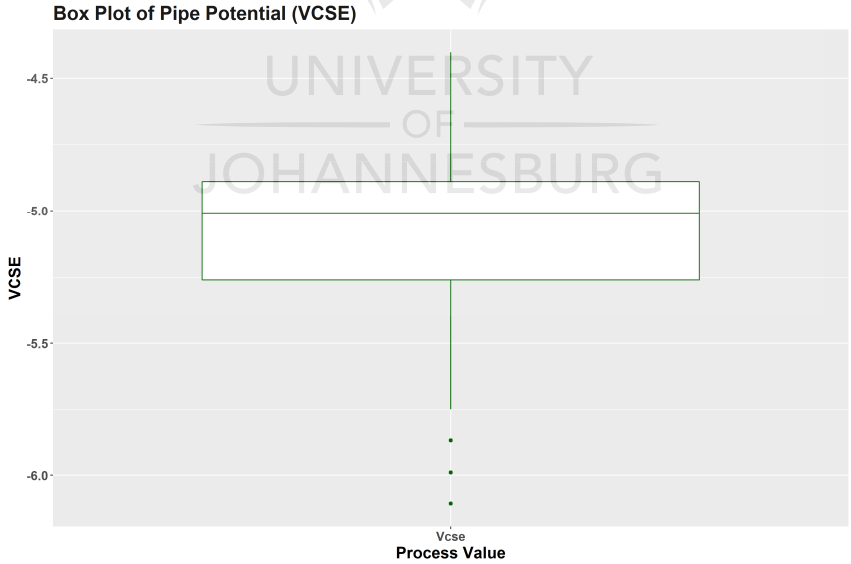

Figure 5-3 - Boxplot of TRU CP pipe potential

From the boxplot above, the CP pipe potential fluctuates between -4.8V_{CSE} and -5.25V_{CSE} (inter-quantile range), with a median value of -5.0V_{CSE}. A lower and upper whisker is present at -4.3V_{CSE} and -6.1V_{CSE}. This boxplot indicates that the CP pipe potential regulated between -4.8V_{CSE} and -5.25V_{CSE}.

Removal of the outliers and centring and scaling the data resulted in minimal improvement in PV correlation:

TRU Outlier-Correlation Comparison			
Dataset	**Variables**	**Correlation**	**% Change**
Original	Vcse vs IOut	-0.6834396	-0.15425%
Original	Vcse vs VOut	-0.6741833	-0.11325%
Original	VOut vs IOut	0.9981833	-0.00125%
Removed Outliers	Vcse vs IOut	-0.6844954	
Removed Outliers	Vcse vs VOut	-0.6749477	
Removed Outliers	VOut vs IOut	0.9981958	
Centred and Scaled	Vcse vs IOut	-0.6844954	0.00000%
Centred and Scaled	Vcse vs VOut	-0.6749477	0.00000%
Centred and Scaled	VOut vs IOut	0.998195	0.00008%

Table 5-2 - TRU Outlier - Correlation Comparison

The skewness had a 3% improvement for the CP pipe potential with outliers removed:

TRU Outlier-Skewness Comparison			
Dataset	**Variables**	**Skewness**	**% Change**
Original	Vcse	-0.4277091	3.99569%
Original	VOut	0.418736	-0.35711%
Original	IOut	0.424159	-0.34900%
Removed Outliers	Vcse	-0.4112758	
Removed Outliers	VOut	0.4202367	
Removed Outliers	IOut	0.4256445	
Centred and Scaled	Vcse	-0.4112758	0.00000%
Centred and Scaled	VOut	0.4202367	0.00000%
Centred and Scaled	IOut	0.4256445	0.00000%

Table 5-3 - TRU Outlier - Skewness Comparison

To determine if the removal of the outliers and centring and scaling improves the model accuracy, the RMSE will be evaluated in the sections following.

5.2.2.1.1.2. Creation of ML Datasets

Three datasets were created, namely the training and test sets, with a ratio of 40%:60% respectively. Furthermore, to ensure reproducibility of the results in R, the seed was set to 1 and sample kind to “Rounding”.

The training dataset is referred to as *train_set*, while the test dataset as *test_set*. The variable names referenced in the sections below are as follows:

Variable Names for ML Models		
Variable	**Description**	**Units**
Vcse	Instant-on CP pipe potential	V_{CSE}
IOut	Rectifier Output Current	A
VOut	Rectifier Output Voltage	V
Idrain	Rectifier Drainage Current	A

Table 5-4 - Variable Names for ML Models

5.2.2.1.1.3. Prediction and Learning

The first prediction focussed on predicting the CP pipe potential. A multiple linear regression (LR) model was trained/fitted using the train_set and with IOut and VOut used as the predictor variables :

```
Call:
lm(formula = Vcse ~ IOut + VOut, data = train_set)

Residuals:
     Min       1Q   Median       3Q      Max
-1.61524 -0.11312 -0.00794  0.12491  0.76629

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -5.062365   0.006197 -816.93   <2e-16 ***
IOut        -0.145384   0.003864  -37.63   <2e-16 ***
VOut         0.052790   0.001857   28.43   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1863 on 25053 degrees of freedom
Multiple R-squared:  0.4836,    Adjusted R-squared:  0.4835
F-statistic: 1.173e+04 on 2 and 25053 DF,  p-value: < 2.2e-16
```

Figure 5-4 - Summary of LM Model (Steady-State TRU)

This formula describes the regression model:

$$V_{CSE} = -5.06236 - 0.14538\text{IOut} + 0.05279\text{VOut}$$

The P-values for the IOut and VOut predictors are statistically significant due to the presence of the * symbols, and the P-value is very small, whereas the absolute t-values for IOut and VOut is more than 2, which indicates high confidence when used as predictors. The R-squared result evaluates the variance described in the model, while the F-statistic determines if a variable’s weight is larger than 0.

The Durbin-Watson [111] test result was 1.994, which indicates an alternative hypothesis with a variable auto-correlation more than 0. The Jarque-Bera test evaluates the normality of the distribution [112], which provided a P-Value of fewer than 2.2 e-16, indicating a skewed/not normal distribution.

To evaluate the prediction accuracy, the predict function was executed in R against the test_set. The RMSE, R^2 and MAE results are tabled below:

Basic Multiple Linear Regression Results		
RMSE	**R^2**	**MAE**
0.1857401	0.4852551	0.1504428

Table 5-5 - Basic Multiple LR Results (Steady-State TRU)

The RMSE indicates a prediction error of 0.18 $V_{CSE,}$ which translates to an absolute prediction error rate of 3.69% and accuracy of 96.30%. A visual representation of the linear regression formula and the variable relationship was calculated in Microsoft Excel, and a line graph was created:

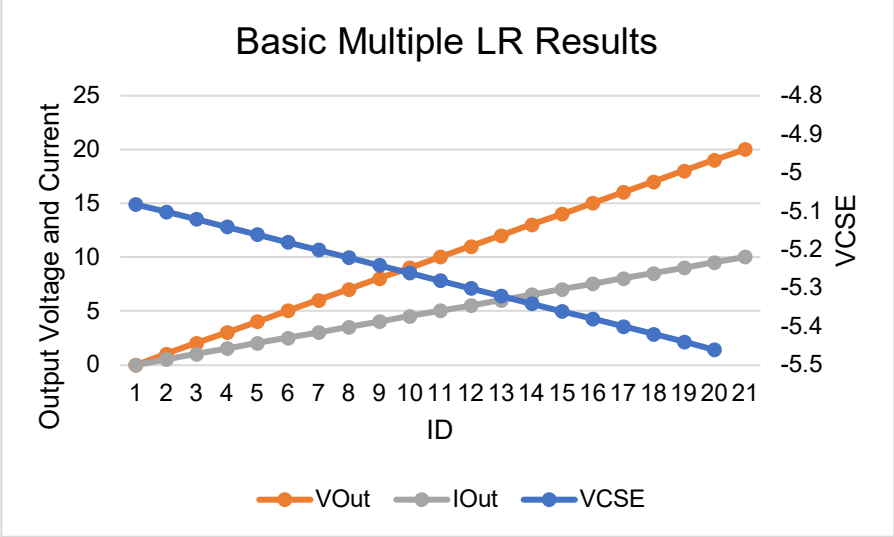

Figure 5-5 - Basic Multiple LR Evaluation Line Graph (Steady-State TRU)

In an attempt to improve the prediction accuracy, a further evaluation of different ML techniques is required. The literature review evaluated some ML techniques applicable to regression and classification algorithms.

The caret (Classification And Regression Training) package [113] was selected since it contains various techniques based on the predictive modelling application. A full list of the models available in the caret package is tabled in Appendix D2.

For this study's regression and classification analysis, the following models were selected for evaluation:

Model Selection Criteria		
Model	**Primary Functions**	**Motivation for Selection**
Linear Regression Model (lm)	Linear Regression (basic and multiple).	The preliminary model for linear regression analysis using Ordinary Least Squares (OLS) for reducing residuals. The basis for determining if linear regression is feasible.
Robust Linear Model (rlm)	Linear Regression, with case weights.	Uses case weights whereby points are not treated equally in an attempt to reduce residuals (where errors are not a normal distribution).
Generalized Linear Model (glm)	Classification and Linear Regression	The basic model for linear regression problems. The basis for determining if linear regression is feasible.
Generalized Additive Model using Splines (gam)	Classification, Linear Regression, Additive and Generalized LR	Improvement of the GLM model by smoothing the data to reduce residuals.
Generalized Additive Model using LOESS (gamLoess)	Classification, Linear Regression, Additive and Generalized LR	Enables smoothing and robustification of outliers to reduce residuals.

Model Selection Criteria		
Model	**Primary Functions**	**Motivation for Selection**
Boosted Generalized Linear Model (glmboost)	Classification, Linear Regression, Boosting, Ensemble Models, Linear Classifier	The boosting process adds the next model to the current model to improve the accuracy of the model.
Boosted Linear Model (BstLm)	Classification, Linear Regression, Boosting, Ensemble Models, Implicit Feature Selection	Boosting improves model accuracy by determining the number of iterations to maximize the likelihood or pseudo-R^2.
Support Vector Machines with Linear Kernel (svmLinear)	Classification, Linear Regression, Kernel Method, Ensemble Models, Linear Classifier, Support Vector Machines	The SVM algorithm classifies data in different hyperplanes in an attempt to improve the prediction accuracy.
Random Forest (rf)	Classification, Linear Regression, Bagging, Ensemble Models, Implicit Feature Selection	Create multiple decision-trees and average deep decision-trees to prevent overfitting of individual trees.
Neural Network (nnet)	Classification, Linear Regression, Case Weights, Neural Network	Creates an artificial neural network to predict classifier labels.

Table 5-6 - Model Selection Criteria - Adapted from Sources [113], [114]

The relevant models were evaluated and presented the following results:

Various Models Evaluation Results			
Dataset	**Model**	**% Error**	**RMSE**
Train/Test	glm	2.980%	0.186
Train/Test	gamLoess	2.897%	0.181
Train/Test	gam	2.832%	0.177
Train/Test	rf	2.270%	0.153
Train/Test	lm	2.980%	0.186
Train/Test	rlm	2.967%	0.186
Train/Test	glmboost	3.091%	0.189
Train/Test	BstLm	3.496%	0.218
Train/Test	svmLinear	2.940%	0.188

Table 5-7 - Various ML Model Evaluation Results (Steady-State TRU)

From the results above, the untuned RF, gam and gamLoess models provided the best RMSE results (prediction error less than 3%), with the RF model performing the best (RMSE of 0.153).

Analysing the predicted against actual values indicated that most models could predict the trend (with varying accuracy), but predictor combinations not present in the training dataset, results in the model using the mean value:

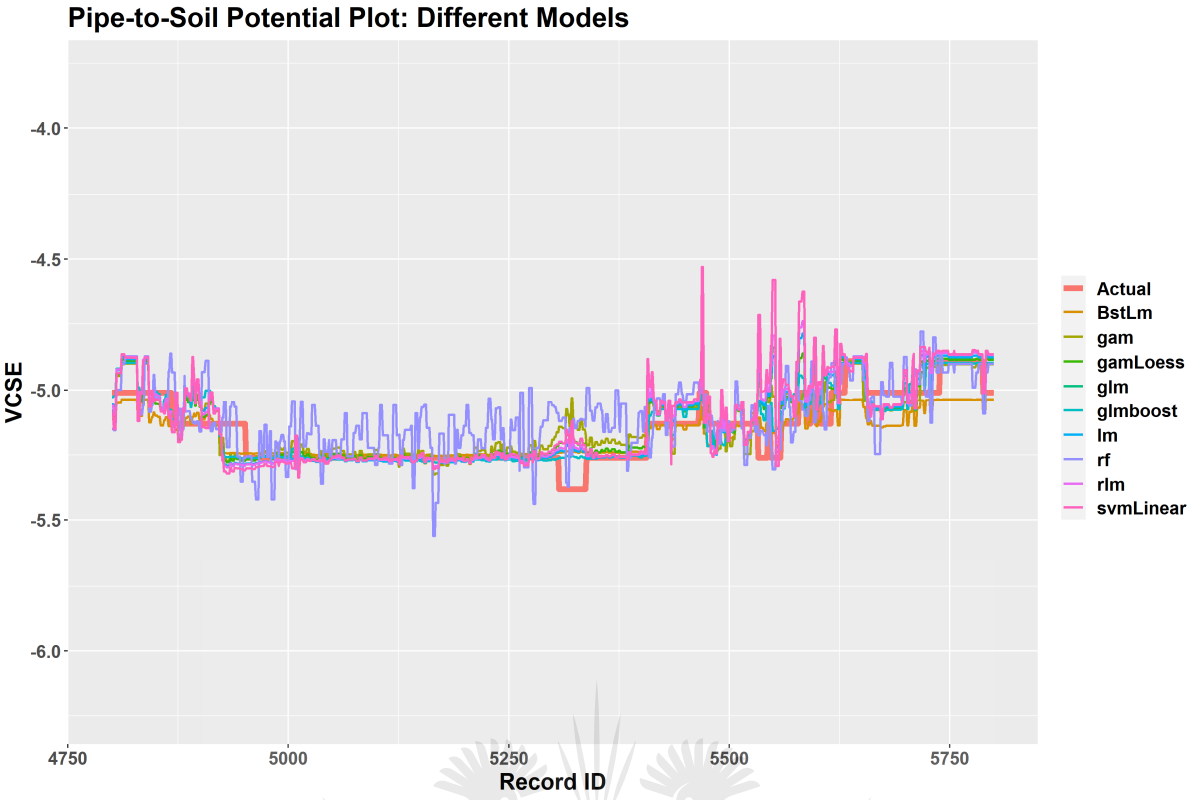

Figure 5-6 - Various Models Line Graphs (Steady-State TRU)

5.2.2.2. Stray Current Operation

Following the findings of chapter four of an FDU operating with stray current, this section evaluates the CP pipe potential prediction accuracy of an FDU, with stray current interference. The raw data received from the SCADA system has a sampling interval of 30 seconds and a period of 30 days.

Similar to the preceding section, a multiple linear regression model was trained/fitted using the train_set with IOut, VOut and IDrain used as the predictor variables:

```
Call:
lm(formula = Vcse ~ IOut + VOut + IDrain, data = train_set)

Coefficients:
(Intercept)         IOut         VOut       IDrain
   -48.9426       1.3832       0.5567       0.2591
```

Figure 5-7 - Summary of LM Model (FDU)

This formula describes the regression model:

$$V_{CSE} = -48.9426 + 1.3832\,\mathbf{IOut} + 0.5567\mathbf{VOut} + 0.2591\mathbf{IDrain}$$

The prediction accuracy for the above formula is, however, not very high based on the RMSE of 26.9.

Basic Multiple Linear Regression Results - Malfunctioning FDU		
RMSE	**R2**	**MAE**
26.953303	0.6522186	18.549177

Table 5-8 - Basic Multiple LR Results (FDU Malfunctioning)

Further investigation of the FDU performance, indicated that the FDU was malfunctioning, which results in a high prediction error rate. To re-evaluate the model, a functioning FDU was selected for analysis (with the presence of stray current).

This formula describes the regression model of the operational FDU:

$$V_{CSE} = -3.50961 + 0.02371\mathbf{IOut} - 0.3360\mathbf{VOut} - 0.04588\mathbf{IDrain}$$

The prediction accuracy for the above formula improved to an RMSE of 2.055:

Basic Multiple Linear Regression Results - Stray Current		
RMSE	**R2**	**MAE**
2.0557106	0.9469795	0.9597756

Table 5-9 - Basic Multiple LR Results (Stray Current)

In an attempt to improve the model prediction accuracy, the following models were also evaluated (similar modelling approach as for the previous section):

Various Models Evaluation Results - FDU Stray Current			
Dataset	**Model**	**% Error**	**RMSE**
Train/Test	glm	7.929%	2.056
Train/Test	gamLoess	7.968%	2.050
Train/Test	gam	7.938%	2.051
Train/Test	rf	7.603%	1.926
Train/Test	lm	7.929%	2.056
Train/Test	rlm	8.187%	2.066
Train/Test	glmboost	7.931%	2.056
Train/Test	BstLm	17.135%	3.269
Train/Test	svmLinear	8.137%	2.061

Table 5-10 - Various Model Evaluation Results (Stray Current)

Similar to the TRU results, the untuned-RF model presented the best RMSE of 1.926. This results can potentially be improved by tuning the RF model and performing cross-validation.

To determine if the data period and sampling rate affect the prediction accuracy, the FDU data period was extended from one to three months, and the sampling interval increased from 30 seconds to 2 minutes and 5 minutes respectively.

The resultant RMSE results are as follows:

Basic Multiple Linear Regression Results - Stray Current - 2 Minutes/3 Months			Basic Multiple Linear Regression Results - Stray Current - 5 Minutes/3 Months		
RMSE	**R2**	**MAE**	**RMSE**	**R2**	**MAE**
0.6752808	0.3886836	0.310488	0.701479	0.3696683	0.3101932

Table 5-11 - Evaluation Results (Stray Current and Data Changes)

Both models yielded a better RMSE with more data and increasing the sampling interval from 30 seconds to a minute interval. The model using a 2-minute sampling interval performed better than the model with a 5-minute sampling interval.

5.2.2.3. Pipeline Section

To predict the CP pipe potentials of a pipeline section depends on the availability of data (predictors) for a specific output of the TRU or FDU. Most of the data recorded at TP's are CP pipe potentials only, and prediction of the CP pipe potential will not be possible using the methods discussed above.

Based on the LR regression formula in the preceding sections, the predictor coefficients determines the slope of the curve. Calculating the change in output current coefficient for downstream TP's (based on previous data), the CP pipe potential at TP's can be estimated for a specific TRU or FDU output:

LR Estimation of VCSE at TP 1				
Rectifier Information			**TP1**	
Vout	**Iout**	**VCSE**	**Iout Coef**	**VCSE**
45.08	10.20	-6.67	-1.379	-3.36
44.95	10.17	-6.66	-1.379	-3.37
45.04	10.23	-6.73	-1.379	-3.42
45.04	10.23	-6.73	-1.379	-3.42
44.95	10.17	-6.66	-1.379	-3.37
44.95	10.17	-6.66	-1.379	-3.37
45.01	10.13	-6.57	-1.379	-3.29

Table 5-12 - LR Estimation of V_{CSE} for TP1

Visualizing the above table for a pipeline section yields the following results:

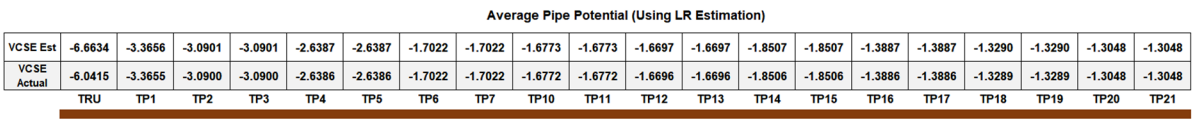

Figure 5-8 - CP pipe potential Comparison for Pipeline Section (Using LR)

The difference between the actual and estimated CP pipe potentials for the TP's are 0.0014 V_{CSE}.

5.2.3. Summary

The first section evaluated various ML models to predict the CP pipe potential of a TRU operating at steady-state. The best prediction accuracy achieved resulted in an RMSE value of 0.153 when using the RF model. When plotting the actual values against the predicted values, it was evident that the model will output an average value, if the predictor combination was not present in the training set.

The second portion of this section evaluated an FDU that malfunctioned and an FDU operating with the presence of stray current. The CP pipe potential prediction error was very high for the malfunctioning FDU (RMSE of 26.9). Evaluating the FDU with stray current interference, presented the best prediction accuracy using the RF model (RMSE of 1.8).

In the next section, the FDU stray current models were evaluated with more training data (3 months), and the sampling interval was increased from 30 seconds to 2 and 5 minutes, respectively. The basic LR model RMSE improved to 0.67 and 0.71, respectively. This improvement is due to the elimination of the CP pipe potential noise, which follows a similar trend than the time-series trend component analysis.

The last section evaluated the estimation of the CP pipe potentials for a pipeline section based on the TRU output current and output voltage through modification of the LR formula for the specific TRU and previous SCADA and logger data.

In conclusion, the linear regression formula describing each unique FDU or TRU, will not be constant and in practice, needs to be determined for each unit individually and the training data period and sampling rate will also be unique for the specific unit.

The next section evaluates the CP pipe potential state prediction using a classification ML approach (using state labels).

5.3. CP pipe potential State Prediction

The preceding results indicated that the prediction accuracy of the CP pipe potential decrease with the presence of stray current. This section evaluates an ML model using a classification approach in R, to predict the Status column value as defined in chapter three.

5.3.1. Status Column Value

The status column assigns a label to a column based on the CP pipe potential's conformance to the defined OW for this study:

- Protected (P) = CP pipe potential is within OW
- Over-protected (OP) = CP pipe potential is more negative than OW
- Under-protected (UP) = CP pipe potential is more positive than OW

5.3.2. ML Performance Evaluation

The RMSE and MAE metrics were used in the previous sections; however, in this section, the performance was evaluated by calculating the mean of all the correct predictions.

5.3.3. Prediction and Learning

The RF, NNET and SVM classification models were evaluated in R for both steady-state and stray current operation, and provided the following results:

State Prediction Results: FDU						
Model	**Predictors**	**Unit Type**	**State**	**Data Period**	**Sampling Rate (Minutes)**	**Accuracy**
RF	VOut + IOut + IDrain	FDU	Steady-state	1 Month	0.5	98.924%
svmLinear	VOut + IOut + IDrain	FDU	Steady-state	1 Month	0.5	98.696%
nnet	VOut + IOut + IDrain	FDU	Steady-state	1 Month	0.5	98.981%
RF	VOut + Iout	FDU	Steady-state	1 Month	0.5	98.787%
svmLinear	VOut + Iout	FDU	Steady-state	1 Month	0.5	98.469%
nnet	VOut + Iout	FDU	Steady-state	1 Month	0.5	98.868%
RF	VOut + IOut + IDrain	FDU	Steady-state	4 Months	2	98.567%
svmLinear	VOut + IOut + IDrain	FDU	Steady-state	4 Months	2	98.669%
nnet	VOut + IOut + IDrain	FDU	Steady-state	4 Months	2	98.694%
RF	VOut + Iout	FDU	Steady-state	4 Months	2	98.426%
svmLinear	VOut + Iout	FDU	Steady-state	4 Months	2	98.669%
nnet	VOut + Iout	FDU	Steady-state	4 Months	2	98.725%
RF	VOut + IOut + IDrain	FDU	Steady-state	4 Months	5	98.608%
svmLinear	VOut + IOut + IDrain	FDU	Steady-state	4 Months	5	98.772%
nnet	VOut + IOut + IDrain	FDU	Steady-state	4 Months	5	98.735%
RF	VOut + Iout	FDU	Steady-state	4 Months	5	98.402%
svmLinear	VOut + Iout	FDU	Steady-state	4 Months	5	98.772%
nnet	VOut + Iout	FDU	Steady-state	4 Months	5	98.772%
RF	VOut + IOut + IDrain	FDU	Stray Current	1 Month	0.5	93.891%
svmLinear	VOut + IOut + IDrain	FDU	Stray Current	1 Month	0.5	88.890%
nnet	VOut + IOut + IDrain	FDU	Stray Current	1 Month	0.5	93.785%
RF	VOut + Iout	FDU	Stray Current	1 Month	0.5	93.656%
svmLinear	VOut + Iout	FDU	Stray Current	1 Month	0.5	88.754%
nnet	VOut + Iout	FDU	Stray Current	1 Month	0.5	93.525%

Table 5-13 - State Prediction Results

Based on the results above, the accuracy of the three models is very high (>98%) for any combination of predictors, data period and sampling rate when no stray current is present. With the presence of stray current, the accuracy drops to a range between 88% and 93% (dependant on model and predictors). The accuracy however improved in comparison to the models trying to predict the CP pipe potential (83%).

The high-accuracy, however, comes with a trade-off as the three models only predict one of three outcomes and not the CP pipe potential as well. If both are results are required, the candidate suggests running both the LR regression and classification models. The selection of the classification model will depend on the computational overhead of training the models.

5.3.4. Summary

This section evaluated the prediction of state labels using a classification approach and using thee different ML models. The accuracy achieved when operating at steady-state was very high (>99%), but the accuracy for stray current was 93% (which is, however, an improvement of the LR approach).

The next section discusses time-to-state prediction.

5.4. Time-to-State Prediction

Survival analysis was performed in R to evaluate the time-to-state prediction using the Survival package and the cycle times defined in chapter three. The results are described below:

5.4.1. Event Values

The Survival package requires data in the following binary format:

- Event Occurred (Pipe Not Protected) = 1
- No Event (Pipe Protected) = 0

Three additional columns were created that maps the state value (OP, UP or P) to a 0 or 1:

	category	randtime	Timestamp	OP	UP	P	Event	TimePrev
1	UP	1577836800	2020-01-01 02:00:00	0	1	0	1	<NA>
2	UP	1577836830	2020-01-01 02:00:30	0	1	0	1	2020-01-01 02:00:00
3	UP	1577836860	2020-01-01 02:01:00	0	1	0	1	2020-01-01 02:00:30
4	UP	1577836890	2020-01-01 02:01:30	0	1	0	1	2020-01-01 02:01:00
5	UP	1577836920	2020-01-01 02:02:00	0	1	0	1	2020-01-01 02:01:30

Figure 5-9 - Survival Event Columns

5.4.2. Event Times

The event times were calculated by ordering the data according to timestamp and calculating the time difference between rows with the same state. Another column was created that calculates the cumulative event time:

	category	randtime	Timestamp	OP	UP	P	Event	TimePrev	tdiff	tdiffval	cumTime
1	UP	1577836800	2020-01-01 02:00:00	0	1	0	1	0	0 secs	0	0
2	UP	1577836830	2020-01-01 02:00:30	0	1	0	1	2020-01-01 02:00:00	30 secs	30	30
3	UP	1577836860	2020-01-01 02:01:00	0	1	0	1	2020-01-01 02:00:30	30 secs	30	60
4	UP	1577836890	2020-01-01 02:01:30	0	1	0	1	2020-01-01 02:01:00	30 secs	30	90
5	UP	1577836920	2020-01-01 02:02:00	0	1	0	1	2020-01-01 02:01:30	30 secs	30	120

Figure 5-10 - Survival Time Columns

5.4.3. KM Modelling

To perform the time-to-event analysis, a KM survival model was fitted based on the Event value (0 or 1) and the event time.

5.4.4. KM Performance Evaluation

To perform the time-to-event analysis, a KM survival model was fitted based on the Event value (0 or 1) and the event time. Using the median time from the output of the fitted survival model, the survival time is estimated to a probability of 50%. The results

below indicate the survival time or time to a not-protected pipeline (which is more than 6000). For this specific ICCP unit evaluated, the pipeline was protected for the 98% of the analysis (the red *Protected Event* curve stays close to 1 for most of the analysis time).

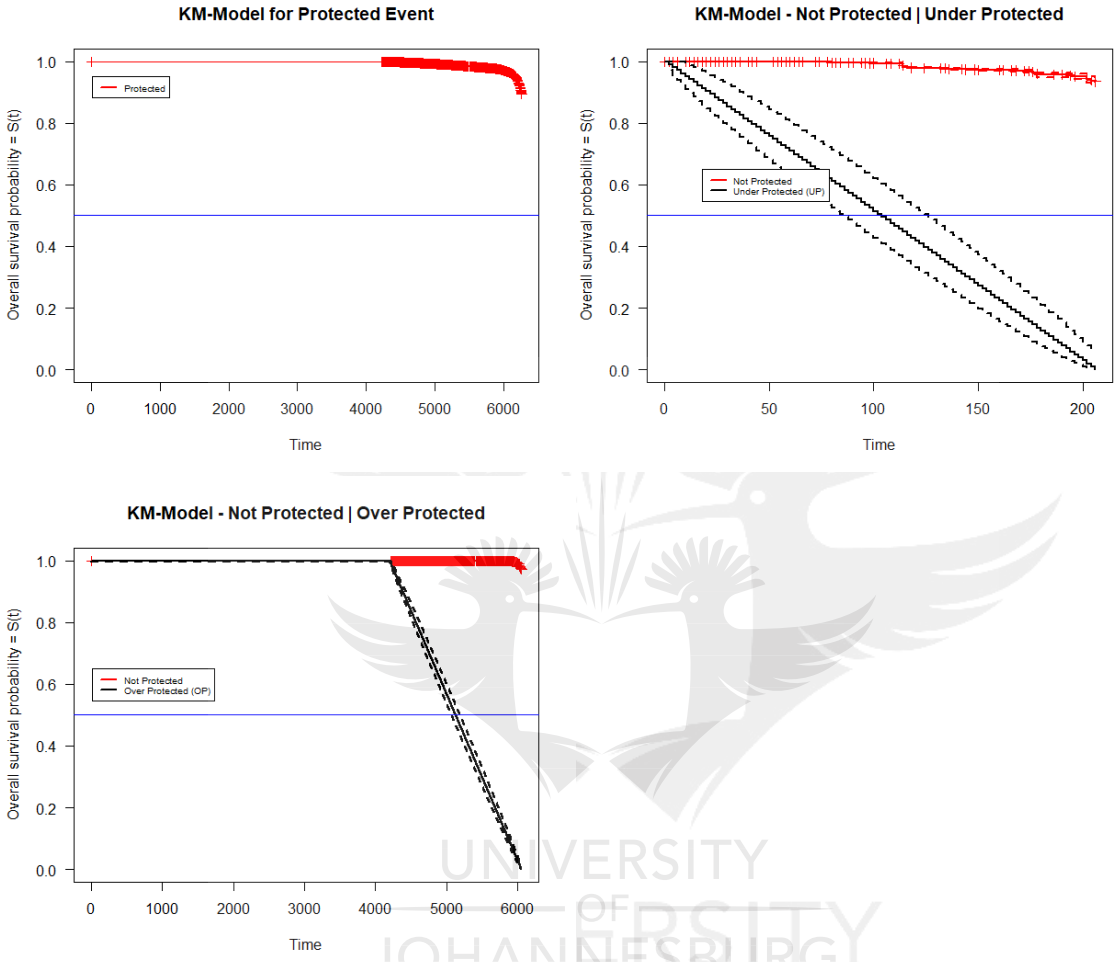

Figure 5-11 - KM Plot for Not-Protected, UP and OP Event

To results are also tabled below:

Time-to-Event Analysis (Times Indicated as minutes)				
	Kaplan-Meier Analysis			
Survival Function	**Event Active (Not Protected)**	**Event Inactive (Not Protected)**	**Event Active (UP=1)**	**Event Active (OP=1)**
Surv(dfS$cumTime, dfS$Event)~ 1	20	5980		
Surv(dfS$cumTime, dfS$Event) ~ dfS$UP			8	
Surv(dfS$cumTime, dfS$Event) ~ dfS$OP				70

Table 5-14 - Time-to-Event Analysis Results

To verify the above results to the actual time the pipeline was not protected, the descriptive time statistics was used for verification. Comparison of the percentage time

of the survival analysis are verified against the percentage statistics of the descriptive statistics yields very close results:

Time-to-Event Analysis								
Kaplan-Meier Analysis			Descriptive Statistics					
% Protected KM	% UP KM	% OP KM	% Protected	% UP	% OP	Protected Time	UP Time	OP Time
99.67%	0.13%	1.17%	98.68%	0.13%	1.19%	29 days 14:12:48	00 days 03:51:41	01 days 10:47:28

Table 5-15 - Time-to-Event Analysis Comparison to Descriptive Statistics

The results from the table above are summarized below:

- The CP pipe potential will shift outside the OW within approximately 29 days (using cumulative time for the 30-day timeframe).
- The CP pipe potential will take approximately 29 days in the 30-day timeframe, to reach a state of UP and will operate in this band for approximately 4 hours.
- The CP pipe potential will take approximately 28 days in the 30-day timeframe, to reach a state of OP and will operate in this band for approximately 1.9 days.

Although this information is useful for historical analysis, a more robust approach is required to estimate the time-to-state change.

5.4.5. Cycle Time Approach

The cycle time approach works on finite pre-defined cycles for which a state can be active (which can typically guide maintenance activities). For this study, two state cycles were defined, namely 40 hours for OP and 24 hours for UP. By decrementing these values on an active state label, the time-to-event, in this case, the cycle end time, can be determined. The table below illustrates this process:

Time-to-Event Analysis using 40-hour Event Cycle			
Event Time	Time Difference	Remaining Cycle Time (Hours)	Estimated Maintenance Time
2020/04/01 01:00:00	**0.00**	**40.00**	**2020/04/02 17:00:00**
2020/04/01 02:00:00	1.00	39.00	2020/04/02 17:00:00
2020/04/01 07:00:00	5.00	34.00	2020/04/02 17:00:00
2020/04/01 14:00:00	7.00	27.00	2020/04/02 17:00:00
2020/04/01 23:00:00	9.00	18.00	2020/04/02 17:00:00
2020/04/02 02:00:00	3.00	15.00	2020/04/02 17:00:00
2020/04/02 06:00:00	4.00	11.00	2020/04/02 17:00:00
2020/04/02 09:00:00	3.00	8.00	2020/04/02 17:00:00
2020/04/02 13:00:00	4.00	4.00	2020/04/02 17:00:00
2020/04/02 17:00:00	**4.00**	**0.00**	**2020/04/02 17:00:00**
2020/04/02 18:00:00	0.00	40.00	2020/04/04 10:00:00
2020/04/02 19:00:00	1.00	39.00	2020/04/04 10:00:00
2020/04/02 20:00:00	1.00	38.00	2020/04/04 10:00:00

Table 5-16 - Time-to-event Results Using 40-Hour Event Cycle

5.4.6. Summary

This section evaluated the time-to-event results based on three defined state labels and time columns (for event time and cumulative time), and KM curves presented the results. Time-to-event and event duration estimation was based on a probability of 0.5 or more of the event occurring. In practice, the probability can be adjusted to evaluate different scenarios.

The time-to-event analysis was also evaluated using a pre-defined cycle time that decrements on event occurrence.

5.5. Maintenance Suggestion

To suggest maintenance activities, the candidate compiled a list of maintenance activities (as defined in chapter three) and created the following additional columns:

- Time Columns – TimePrev, randtime, Time Difference and Cum Time
- Maintenance Suggestion (numeric value as per design) – For OP, UP, stray current, rectifier output and drainage current less than zero

```
           Date     Time  IOut IDrain  VOut  Vcse           Timestamp    ID Status UnitType StatusNum RectOper RiskFactor RiskLevelOP RiskLevelUP
12442 2020-05-14 06:45:00 14.68   2.67 47.08 -2.41 2020-05-14 06:45:00 12442      P        3         1         1          3           0           0
15778 2020-05-25 20:45:00 13.91   3.81 18.93 -3.05 2020-05-25 20:45:00 15778      P        3         1         1          3           0           0
22058 2020-06-16 16:05:00 23.54   3.05 25.86 -4.86 2020-06-16 16:05:00 22058      P        3         1         1          3           0           0
      RiskLevelP MaintOP MaintUP           TimePrev   randtime  tdiff tdiffval cumTime StatusPrev StatusNumPrev StateTransition VcsePrev VcseRes
12442          0       0       0 2020-05-14 06:40:00 1589431500 5 mins       5   62205          P             1               0    -2.19    0.22
15778          0       0       0 2020-05-25 20:40:00 1590432300 5 mins       5   78885          P             1               0    -3.05    0.00
22058          1       0       0 2020-06-16 16:00:00 1592316300 5 mins       5  110285          P             1               0    -4.19    0.67
      StrayCurrent NoCurrent NoDrain MaintStrayCurrent MaintNoCurrent MaintNoDrain OPFactor UPFactor PFactor DistanceFactor TotalDistance TotalRiskPos
12442            0         0       0                0              0            0      0.5      0.8    0.25              1             6           19
15778            0         0       0                0              0            0      0.5      0.8    0.25              1             6           19
22058            0         0       0                0              0            0      0.5      0.8    0.25              1             6           19
      TimeFactor IntervalNr CPRiskLevel  CPHealth
12442       0.15          1       0.00 100.00000
15778       0.15          1       0.00 100.00000
22058       0.15          1       0.25  98.68421
```

Figure 5-12 - New Columns in R for Maintenance Suggestion

The prediction results for the RF, NN and SVM classification models with predictors *VOut* + *IOut* + *Idrain* are tabled below:

Maintenance Suggestion - Prediction Accuracy Results for an FDU								
Model	**State**	**Data Period**	**OP**	**UP**	**Stray Current**	**No Rectifier Output Current**	**No Rectifier Drainage Current**	**Average Prediction Accuracy**
RF	Steady-state	4 Months	98.583%	99.872%	99.776%	100.000%	99.798%	**99.606%**
svmLinear	Steady-state	4 Months	98.892%	99.872%	99.808%	99.989%	99.798%	**99.672%**
nnet	Steady-state	4 Months	98.892%	99.872%	99.808%	99.989%	99.798%	**99.672%**
RF	Stray Current	1 Month	91.563%	94.774%	97.458%	99.991%	99.407%	**96.639%**
svmLinear	Stray Current	1 Month	74.249%	94.468%	97.667%	99.900%	94.579%	**92.173%**
nnet	Stray Current	1 Month	85.428%	94.666%	97.668%	99.980%	99.424%	**95.433%**

Table 5-17 - Maintenance Suggestion Prediction Accuracy

Similar to the classification approach for state prediction, the prediction accuracy is reduced for a rectifier operating with stray current. The RF model provided the best prediction accuracy for the two datasets.

A summary of the maintenance suggestion counts are shown below:

Maintenance Suggestions Per Unit and Activity							
Maintenance Activity				FDU (Stray Current)		FDU (Steady State)	
Index	Fault	Risk Level	Required Remedial Action	Total Suggestions for Period	Total Per Category	Total Suggestion for Period	Total Per Category
1	OP	1	Monitor and if required, adjust supplying rectifier	0	21116	0	360
2	OP	2	Adjust supplying rectifier output and monitor performance	6872		304	
3	OP	3	Adjust supplying rectifier output and monitor performance	7844		52	
4	OP	4	Adjust supplying rectifier output, investigate possible interference and monitor	6400		4	
5	UP	1	Monitor and if required, adjust supplying rectifier	0	36744	0	38
6	UP	2	Adjust supplying rectifier output and monitor performance	477		37	
7	UP	3	Adjust supplying rectifier output and monitor performance	148		0	
8	UP	4	Adjust supplying rectifier output, investigate possible interference and monitor	36119		1	
9	Stray Current Interference	1	Investigate causes and adjust rectifier. Initiate interference mitigation projects	1938	1938	56	56
10	Rectifier not supplying current	1	Investigate rectifier and resolve the issue	1318	1318	53	53
11	Rectifier not draining current	1	Investigate rectifier and resolve the issue	4875	4875	2	2

Table 5-18 - Maintenance Suggestions per Unit and Activity

An FDU operating at steady-state requires less maintenance than compared to an FDU operating with stray current. A cycle time (discussed in the preceding section), can estimate the maintenance activity and time.

5.5.1. Summary

This section evaluated the prediction of maintenance activities for an FDU with stray current and at steady-state using the RF, NN and SVM models. The RF models provided the best results for suggesting maintenance activities.

5.6. Conclusion

This chapter evaluated the predictive modelling results in an attempt to answer the research questions and achieve the objectives of this study. Appendix D contains a list of all the R packages used in this study and the models available in the caret package.

The first section evaluated the predictive modelling results for various ML models when predicting the CP pipe potential. The data was prepared by removing outliers and centring and scaling the data. This process, however, did not result in an improvement in the kurtosis, skewness or correlation of data.

Following the data preparation steps, the first objective was to obtain an LR function for a TRU by modelling different predictor combinations. After that, various other LR models was evaluated using the caret package and evaluated using the RMSE metric. The RF model performed the best, yielding the smallest RMSE value of 0.153 or percentage error of 2.27%. The next section focussed on analysing an FDU operating with severe stray current and the RMSE was poor (26.953303); however, evaluation of an FDU operating at steady-state yielded a good RMSE of 1.926. By increasing the data interval and sampling rate, the RMSE reduced to 0.675.

The results indicated two crucial findings; namely, the prediction is accurate if a TRU or FDU is operating within the defined OW. The RF models presented the best RMSE for both a TRU and FDU. However, when a TRU or FDU with stray current is evaluated, the prediction accuracy decreases substantially due to the high variation in data and combinations that might not be present in the training data set.

By determining the LR formula for a TRU on a pipeline section (operating at steady-state), the output current coefficient was calculated based on historical data for each of the TP's along the pipeline. The CP pipe potential was estimated for the TP's based on the estimated output current coefficient.

In an attempt to improve the RMSE of an FDU or TRU operating with stray, a classification ML approach was evaluated by creating three state labels, based on the CP pipe potential (P, OP, UP). Three models were evaluated; namely, RF, NN and SVM and the prediction accuracy increased to 93% for the RF and NN models.

The maintenance activity suggestion evaluation yielded a classification model accuracy of 99% for an FDU and 96% for an FDU operating with the stray current. The time component of the maintenance suggestion was done through the time-to-state analysis.

The time-to-event analysis evaluated the state duration for the three states (OP, UP, P) by calculating the event time and the cumulative time. The test results indicated the estimated time-to-event and the event duration (for the data period) using the survival

package in R. Using a cycle time; the time-to-event can be estimated for events or typical maintenance activities. Analysis of the trend component can inform long term maintenance activities.

Lastly, the maintenance activities were suggested based on the risk level computed for each row, and the results indicated that this approach is a good foundation for suggesting maintenance activities based on historical data.

The next chapter discusses the research results against the research objectives and questions.

6. CHAPTER 6: CONCLUSION AND RECOMMENDATIONS

6.1. Introduction

This chapter reviews the research results and compares the results to the objectives of this study.

This chapter consists of the following sections:

i. Evaluation of research findings and conclusions
ii. Discuss the limitations of this study
iii. Recommendations for future research

6.2. Research Findings

The research objectives were defined as:

1. Determine if statistical analysis of CP data, based on the NACE SP0169-2013 criteria for CP evaluation, can predict or estimate the ICCP unit and downstream TP state.
2. Determine if a maintenance activity can be suggested based on the ICCP unit state.

The research questions were defined as:

1. Which statistical analysis methods can be used on historical and real-time CP data to predict or estimate the state of ICCP units or TP's?
2. Which maintenance activities are required to remedy the ICCP unit state, and what mechanism can be used for suggesting maintenance activities?

The sections below discuss the research findings.

6.2.1. Research Objective 1 – ICCP and TP State Prediction

The literature review investigated the basics of corrosion intending to establish an understanding of corrosion prevention techniques. In particular, the evaluation of CP system operation, design and maintenance, provided the foundation for the data analysis portion of this study. It was imperative to establish which standards and statutes govern the operation of a pipeline network as these standards provides frameworks for evaluating and monitoring of CP systems. The NACE SP0169-2013 standard formed the basis to define an OW for instant-on CP pipe potentials for logical data analysis. The PIMS and CMS review, also indicated that a pipeline CP evaluation should consider past performance.

The literature review also investigated theoretical concepts of reliability engineering principles, in particular, CBM systems and maintenance strategies to build a framework for the data analysis design and evaluation. Seeing that this study focusses primarily on data analytics, it was necessary to determine if a data-driven approach or model-driven approach will suffice. The resultant approach for this study was data-driven as determined by the data exploration and ML model evaluation. The literature review did not present literature applicable to the scope of this study, and several case studies presented ML techniques employed for other CBM or PdM systems.

The first research question is concerned with an evaluation of statistical methods to enable state prediction for ICCP units and related downstream TP's. Data preparation, feature engineering and feature extraction were necessary to enable the evaluation of the research objective of predicting the state on an ICCP unit and downstream TP's.

Since the data available for this study consisted of historical operating data of ICCP units and TP's, additional columns were required that facilitates the data analysis process. The following additional columns were created in the dataset:

- Timestamp – Combination of date and time column
- Index – Increasing numeric row index
- Unit type – CP equipment identification
- Event Time – Determines the event time per status
- Cumulative – Determine the cumulative event time per status
- Status – Determine the row state based on the defined OW (guided by the NACE SP0169-2013 criteria). Three state labels exist, namely OP, P and UP.
- Rectifier Operational – Determined by the output voltage and current as well as the state
- CP Current Spread – Based on TP distance from FDU or TRU
- Rectifier Risk Level – Based on unit criticality for a TP, TRU or FDU
- CP pipe potential Risk Columns – Risk levels assigned as per pre-defined OW's within the OP, P and UP states
- Stray Current Risk Column – Determines the stray current between two data points
- CP Risk Indicator – Formula based on unit type, unit risk level, OP, P and UP risk levels and predefined factors and the unit location. Two formulas were evaluated, namely, one with a proportional change and one with an inverse change for downstream TP's
- CP Health Indicator – Risk indicator displayed as a percentage value

The main starting point for any predictive modelling design or study is to evaluate the composition of the data that will be used for analysis [81]. Applicable to both research objectives, was the data exploration process, to investigate the data fields, trends, distributions and time-series components. Various methods were employed for data exploration to seek information that can inform the modelling process and also relates to question one of this study, namely, identify and evaluate statistical methods for state prediction. Appendix C1 presents a summary of the data exploration techniques used, with their aim and the research objective and questions addressed.

To answer the first research question, the table below identifies the statistical methods to predict the ICCP unit and downstream TP state. Furthermore, the performance results also indicate the feasibility of each prediction based on the specific performance metric. For a complete list of models evaluated, please refer to Appendix C2.

When evaluating predictive models, a performance metric is required to determine the prediction accuracy [81]. For the linear regression models, the RMSE indicated the

estimated prediction error (which translates to an error in predicting the CP pipe potential). The best result was an RMSE of 0.153 when a TRU is operating at steady-state. Similarly, the best RMSE for an FDU operating with stray current was 1.926; however, if the data period and interval are increased, the RMSE decreased to 0.675. Also evident in the modelling was that a malfunctioning FDU had a very high RMSE (26.95), which indicates that the model did not have enough training data to predict the malfunction. The high RMSE can, however, be remedied by increasing the training data or can be used as an indicator for a malfunctioning state.

In an attempt to improve the prediction RMSE, a classification approach consisted of LAD by assigning state labels (Status column) based on the CP pipe potential conformance to the defined OW (P, OP and UP). Overall, the RF model accuracy was improved since the noise was removed from the raw data. For a TRU, the best prediction accuracy was 98.87% and an FDU operating with stray current, 93.66%.

The second part of this objective was to estimate the downstream TP state, based on the ICCP unit state. For this estimation, an LR equation for the supplying TRU was determined, and the output current coefficients were determined for each TP using historical CP pipe potential data (seeing that the current shifts the CP pipe potential). The estimated CP pipe potentials were 99.95% accurate for the TP's.

A CP heath indicator was also evaluated for a pipeline section and presented an average accuracy of 3%, although the accuracy was not uniform across the pipeline evaluated. The estimation process consisted of two scenarios, one where the downstream TP health is estimated based on the ICCP unit output and secondly, where the health was based on the actual data from the ICCP unit and the TP's. Although the accuracy was not linear across the pipeline section, this is a dynamic indicator that can be used with real-time data, if the governing formula's constant factors are adjusted per TP on the line at regular intervals. Descriptive statistics were also computed to determine the CP pipe potential's conformance to the NACE SP0169-2013 criteria over a period (as % or time statistics). These statistics can be used in conjunction with the health indicator to verify conformance.

From the analysis, it was evident that the CP pipe potential can vary based on a change in CP current or due to stray current. Where stray current is present, the variable correlation change and the predictive model becomes inaccurate. Furthermore, the use of continuous and periodic data was evaluated and indicated that a pipe profile could be established using periodic data; however, continuous data is preferred since the sample size can improve the prediction or estimation accuracy.

In conclusion, this study proved that the state of the ICCP unit and downstream TP's could be predicted or estimated if three factors are considered, namely, an OW is defined (either as a numeric value or classification label); the predictive modelling considers historical data; and the constants or coefficients are determined for each individual ICCP unit or TP. Various methods were presented in this study for state prediction and estimation.

The next section discusses the research findings related to objective two.

6.2.2. Research Objective 2 – Suggest Maintenance Activity based on ICCP Unit State

Objective two is concerned with the maintenance activity suggestion based on the ICCP unit state. It was imperative to review the literature available on maintenance strategies, CM and the implementation of PdM systems. Furthermore, consultation of the NACE standard and CFR statute provided a foundation of the absolute requirements with regards to CP system maintenance. Both the CFR and NACE standards suggested a time-based maintenance approach that stipulates the inspection frequencies of different CP equipment. For a small pipeline, a time-based maintenance approach is economically feasible; however, for more extensive pipeline networks, the risk increases if the CP system is malfunctioning and hence impacts the maintenance cost.

The research question related to this objective indicated that two primary activities were required, namely, the compilation of a maintenance matrix based on various ICCP conditions, and the statistical evaluation of suggested maintenance activities.

Based on the literature review, maintenance approaches can include a combination of risk-based, state-based and time-based maintenance. The maintenance approach for this study considered a combination of the mentioned maintenance approaches (risk, time and state). A maintenance activity matrix was developed that included the states and risk factors defined in the feature engineering section of this study. The time section was based on an estimate of how long a specific condition is allowed to occur.

The developed maintenance matrix for this study answer the research question as to which maintenance activities are required. The activities considered the states (P, OP and UP), as well as the associated risk and a selected time window. Furthermore, remedial action was suggested by the candidate based on various literature and industry experience.

The time window for maintenance is significant for the context of this study, seeing that it can not compare to a system that has an abrupt fault (example a quick pressure drop or a circuit-breaker trip). The time window was defined based on the impact of either over-protecting (OP) or forcing corrosion (UP) on the pipeline. In practice, this time window should consider the pipeline wrapping condition as well as the corrosion and should be adjusted accordingly.

The maintenance matrix also included the following faults:

- Stray Current – Voltage variance exceeding setpoint (NACE defines this setpoint as a 20%)
- Rectifier not supplying current
- Rectifier not draining current (FDU only)
- Pipe AC potential exceeding 15VAC

To suggest a maintenance activity, and answering the second part of the research question, a risk and event column was created based on the state (P, OP and UP). An ML classification approach then suggested the required maintenance activity with a

99% prediction accuracy using the RF model. This model did, however, not consider the time-duration of an event.

Three approaches were considered to determine the maintenance time components. The first approach consisted of modelling the survival analysis in R Studio based on the KM-curve and estimating the event time based on a probability setpoint of 0.5. This approach considered the historical data and can be used to suggest a maintenance activity if a retrospective analysis is performed. The probability setpoint needs to be fine-tuned per fault condition.

The second approach consisted of assigning a cycle time to each fault (as per the maintenance matrix) and using real-time data that decrements the cycle time on a fault condition and duration and hence forecasting the maintenance time and date. The cycle time resets when reaching zero, and the process repeats. In practice, the cycle time should only reset, once maintenance has been performed.

The third approach consisted of decomposition of a time-series object and evaluating the CP pipe potential trend component over different periods. This approach is also retrospective but can facilitate long-term maintenance activity decisions (for example, seasonal adjustment of rectifiers). For a complete list of models evaluated, please refer to Appendix C3.

In conclusion, the research objective, i.e. suggesting a maintenance activity for an ICCP unit, is obtainable by compiling a maintenance matrix that considers the ICCP states (P, OP and UP) and using an RF ML model to suggest the required maintenance (based on the fault condition). Furthermore, three methods were evaluated to suggest the maintenance time, namely, the retrospective KM survival analysis, the real-time cycle time analysis and using the trend component of a decomposed time-series object.

The next section mentions the study's limitations.

6.2.3. Study Limitations

The study limitations include the following:

i. The scope of this study is limited to ICCP units and TP's.
ii. The CP pipe potential data is instant-on, and no IR-drop was available in the received data sets. The IR-drop was omitted for this study since the RMSE of the ML models will not change if an IR-drop is included.
iii. Only a short pipeline section was evaluated to estimate the TP state.

6.2.4. Recommendations

The predictive maintenance framework can predict either the CP pipe potential or state based on the past performance of an ICCP unit. In practice, the predicted state or CP pipe potential can inform maintenance activities considering both the risk and duration of the particular state. Furthermore, the ML models can also predict the CP pipe potential if there is an error with the CP pipe potential instrument (using the ICCP unit output voltage and current as predictors). The results indicated that the prediction

accuracy could be improved by changing the sampling rate and time intervals of the ML datasets, and in practice, should be considered for each ICCP unit.

Continuous learning of ICCP unit data is recommended to ensure the prediction accuracy is high for different combinations of data. Continuous ICCP and TP data are recommended (using remote monitoring); however, periodic data can also be used for learning and prediction but will present a trade-off in model accuracy. Selection of the ML model should be based on the accuracy and computational overhead. A high RMSE should not be discarded from the onset when evaluating different ML models (as was seen by a malfunctioning FDU). The high RMSE can potentially be used as a status indicator (high RMSE suggests an ICCP unit fault).

Estimation of downstream TP potentials (based on the linear regression coefficients of the supplying ICCP units), can be used for estimating potentials of a pipeline section and consequently deciding on required maintenance activities. The coefficients are unique for each ICCP unit and require separate modelling of each pipeline section. Where stray current is present, the coefficients need to be determined at shorter intervals.

A maintenance matrix considering the function of the ICCP units can potentially optimize maintenance activities since the maintenance is not based on an on or off state of the ICCP unit. The risk level and time-limit for each maintenance activity should be estimated for each pipeline section (since pipelines can have different asset life expectancies or the pipeline condition can deteriorate at a faster rate than other pipelines). Estimating the time-to-maintenance can be based on a running cycle-time method for each state, or performing trend component analysis of the CP pipe potential time-series data.

6.2.5. Recommendations for Future Research

The primary focus of this study was to model the variation of the CP pipe potential (seeing that this is the critical metric for the NACE SP0169-2013 criteria). For future research, the following is recommended:

i. Expand the ML analysis to include more rectifier monitoring points, such as the output frequency, the resistance of the anode-bed, the coupon current and voltage and environmental monitoring (temperature and humidity). Resistance probes can also aid in predicting the corrosion rate.
ii. Expand the ML analysis to include ACM stations, cross-bonds and results from other measurements (CIPS, DCVG, ACVG and PCM).
iii. Improve the prediction accuracy of an FDU or TRU operating with stray current.
iv. Extend the CP health monitoring for longer pipeline sections.
v. Evaluate a model-driven approach against this data-driven approach.
vi. Extend the modelling to include fault-pattern identification of ICCP units.

6.2.6. Conclusion

In the quest to reduce the cost associated with maintenance and corrosion of pipelines, this study was conducted to establish and evaluate a predictive maintenance framework for ICCP units based on the CP pipe potential. The CP pipe

potential is significant, as it needs to conform to any of the three criteria specified in the NACE SP0169-2013 standard.

By using historical data collected from a CP SCADA system and logger recordings, the candidate was able to predict the CP pipe potential, the state (OP, UP and P) and suggest maintenance activities (as per a defined maintenance matrix).

The prediction results indicate that a predictive maintenance approach, based on the states (OP, UP and P) is feasible and can potentially reduce the maintenance cost associated with extensive pipeline networks.

7. REFERENCES

[1] P. Sun, L. Wang, W. Zhu, and J. Wang, “Statistical Analysis of Underground Pipeline Typical Accidents and Numerical Simulation Research,” in *Proceedings - 2nd International Conference on Data Science and Business Analytics, ICDSBA 2018*, 2018, doi: 10.1109/ICDSBA.2018.00-39.

[2] P. Lochner, S. Laurie, S. Bundy, and C. Green, “Strategic Environmental Assessment for the Development of a Phased Gas Pipeline Network in South Africa FINAL SEA REPORT.”

[3] C. Geck, “The World Factbook,” *Charlest. Advis.*, vol. 19, no. 1, pp. 58–60, Jul. 2017, doi: 10.5260/chara.19.1.58.

[4] L. Njomane and A. Telukdarie, “Corrosion management: A case study on South African oil and gas company,” in *Proceedings of the International Conference on Industrial Engineering and Operations Management*, 2018.

[5] C. I. Ossai, “Advances in Asset Management Techniques: An Overview of Corrosion Mechanisms and Mitigation Strategies for Oil and Gas Pipelines,” *ISRN Corros.*, 2012, doi: 10.5402/2012/570143.

[6] G. Koch, J. Varney, N. Thopson, O. Moghissi, M. Gould, and J. Payer, “International Measures of Prevention , Application , and Economics of Corrosion Technologies Study,” 2016.

[7] E. G. P. Marshall E. Parker, “Fundamentals of Corrosion,” in *Pipeline Corrosion and Cathodic Protection*, 3rd ed., Elsevier, 1999, pp. 146–149.

[8] K. Datta and D. R. Fraser, “A corrosion risk assessment model for underground piping,” in *Proceedings - Annual Reliability and Maintainability Symposium*, 2009, doi: 10.1109/RAMS.2009.4914686.

[9] ConstructionMentor, “Backfilling Pipe Trenches,” 2020. [Online]. Available: https://constructionmentor.net/backfilling-pipe-trenches/. [Accessed: 02-May-2020].

[10] E. G. P. Marshall E. Parker, “Cathodic Protection of Steel in Soil,” in *Pipeline Corrosion and Cathodic Protection*, 3rd ed., E. G. P. Marshall E. Parker, Ed. Elsevier, 1999, pp. 150–152.

[11] NACE International, *NACE Standard SP0169 - Standard Practice Control of External Corrosion on Underground or Submerged Metallic Piping Systems*. 2013.

[12] R. Kumar and M. L. Dewal, “Multi-Supervisory Control and Data Display,” *Int. J. Comput. Appl.*, vol. 2, no. 1, pp. 1–5, 2010, doi: 10.5120/619-870.

[13] T. A. Runkler, *Data Analytics*. Wiesbaden: Springer Fachmedien Wiesbaden, 2016.

[14] O. Motaghare, A. S. Pillai, and K. I. Ramachandran, “Predictive Maintenance Architecture,” in *2018 IEEE International Conference on Computational Intelligence and Computing Research, ICCIC 2018*, 2018, doi:

10.1109/ICCIC.2018.8782406.

[15] M. Kuhn *et al.*, “Classification and Regression Training,” *CRAN*, 2020. [Online]. Available: https://cran.r-project.org/web/packages/caret/caret.pdf. [Accessed: 03-Oct-2020].

[16] C. L. Winchester and M. Salji, “Writing a literature review,” *J. Clin. Urol.*, vol. 9, no. 5, pp. 308–312, Sep. 2016, doi: 10.1177/2051415816650133.

[17] H. Snyder, “Literature review as a research methodology: An overview and guidelines,” *J. Bus. Res.*, 2019, doi: 10.1016/j.jbusres.2019.07.039.

[18] A.W. PEABODY, *CONTROL OF PIPELINE CORROSION*, 2nd ed. Houston: NACE International, 2001.

[19] U. Anthony, M. Ikenna, O. B. Ufuma, and E. Dt, “Corrosion Rates and its Impact on Mild Steel in Some Selected Environments,” *J. Sci. Eng. Res.*, 2016.

[20] T. F. Lewicki and N. L. Fowler, “The effect of corrosion myths on national electrical standards,” in *IEEE Conference Record of Industrial and Commercial Power Systems Technical Conference*, 1992, doi: 10.1109/icps.1992.163383.

[21] R. W. Revie and H. H. Uhlig, *Corrosion and Corrosion Control: An Introduction to Corrosion Science and Engineering: Fourth Edition*. 2008.

[22] W. D. Callister, “Materials science and engineering: An introduction (2nd edition),” *Mater. Des.*, 1991, doi: 10.1016/0261-3069(91)90101-9.

[23] J. P. Guyer, “Introduction To Cathodic Protection.,” *Pipes Pipelines Int.*, 2009.

[24] ASTM, *ASTM G96-90 (2018): Standard Guide for On-Line Monitoring of Corrosion in Plant Equipment (Electrical and Electrochemical Methods)*. 2018.

[25] G. Inzelt, A. Lewenstam, and F. Scholz, *Handbook of reference electrodes*. 2013.

[26] L. Yang, *Techniques for corrosion monitoring*. 2008.

[27] M. G. Fontana, N. D. Greene, and J. Klerer, “Corrosion Engineering,” *J. Electrochem. Soc.*, 1968, doi: 10.1149/1.2411256.

[28] NACE International, *NACE CP 2 – Cathodic Protection Technician Training Manual*. NACE International, 2013.

[29] R. M. Park, “A guide to understanding reference electrode readings,” *Mater. Perform.*, 2009.

[30] S. Szabó and I. Bakos, “Reference electrodes in metal corrosion,” *International Journal of Corrosion*. 2010, doi: 10.1155/2010/756950.

[31] F. J. Ansuini and J. R. Dimond, “Factors affecting the accuracy of reference electrodes,” *Mater. Perform.*, 2017.

[32] E. McCafferty, *Introduction to corrosion science*. 2010.

[33] NACE International, *NACE CP 3 – Cathodic Protection Technologist Training Manual*. NACE International, 2005.

[34] J. R. Dimond and F. J. Ansuini, “Effect of measurement and instrumentation errors on potential readings,” in *NACE - International Corrosion Conference Series*, 2001.

[35] NACE International, *NACE CP 1 – Cathodic Protection Tester Training Manual*. 2008.

[36] ASM-13A, “Vol 13A - Corrosion: Fundamentals, Testing, and Protection,” *ASM Handb.*, 2003.

[37] S. M. Bashi, N. F. Mailah, and M. A. Mohd Radzi, “Cathodic protection system,” in *Proceedings. National Power Engineering Conference, 2003. PECon 2003.*, pp. 366–370, doi: 10.1109/PECON.2003.1437476.

[38] L. C. Wrobel and P. Miltiadou, “Genetic algorithms for inverse cathodic protection problems,” *Eng. Anal. Bound. Elem.*, vol. 28, no. 3, pp. 267–277, Mar. 2004, doi: 10.1016/S0955-7997(03)00057-2.

[39] C. Engineering, “Cathodic Protection Rectifier & Electrical Systems,” 2019. [Online]. Available: https://www.cathtect.com/cathodic-protection-rectifiers-electrical-systems/. [Accessed: 06-Sep-2020].

[40] GCP German Cathodic Protection GmbH, “German Cathodic Protection,” 2020. [Online]. Available: http://www.gcp.de/index.php/en/datasheets.html. [Accessed: 06-Sep-2020].

[41] B. Martin, “CATHODIC PROTECTION IN THE WATER INDUSTRY | TECHNICAL EVALUATION,” *Brian Martin & Associates*, 2020. [Online]. Available: https://membership.corrosion.com.au/blog/cathodic-protection-in-the-water-industry-technical-evaluation/. [Accessed: 04-Sep-2020].

[42] U.S. Government Publishing Office, “Code of Federal Regulations,” *Food Standards and Definitions in the United States*, 1963. [Online]. Available: https://www.govinfo.gov/app/collection/cfr. [Accessed: 15-Aug-2020].

[43] PHMSA, “PIPELINE SAFETY REGULATIONS - 49 CFR PART 192.” 2015, Oklahoma City, pp. 60–85, 2015.

[44] W. Brian Holtsbaum, *Cathodic Protection Survey Procedures*, 2nd ed. Houston: NACE International, 2012.

[45] L. D. Bloomberg and M. Volpe, “A Complete Dissertation: The Big Picture, Chapter 1 Objectives,” *Complet. Your Qual. Diss. A Road Map From Begin. to End*, 2017.

[46] NACE International, “NACE Standard TM0497-2018-SG - Measurement Techniques Related to Criteria for Cathodic Protection on Underground or Submerged Metallic Piping Systems.” NACE International, Houston, 2018.

[47] Hart Energy, “100-mv shift best for older lines,” 2002. [Online]. Available: https://www.hartenergy.com/news/100-mv-shift-best-older-lines-51346. [Accessed: 06-Sep-2020].

[48] G. Cui, Z. L. Li, C. Yang, and M. Wang, “The influence of DC stray current on pipeline corrosion,” *Pet. Sci.*, 2016, doi: 10.1007/s12182-015-0064-3.

[49] I. A. Metwally, H. M. Al-Mandhari, A. Gastli, and Z. Nadir, “Factors affecting cathodic-protection interference,” *Eng. Anal. Bound. Elem.*, 2007, doi: 10.1016/j.enganabound.2006.11.003.

[50] K. Vranešić, S. Lakušić, and M. Serdar, “Stray current corrosion activity on rail transit system in urban areas,” in *Road and Rail Infrastructure V*, 2018, doi: 10.5592/co/cetra.2018.936.

[51] NACE International, *NACE Standard RP0104-2004 - The Use of Coupons for Cathodic Protection Monitoring Applications*. Houston, 2004.

[52] E. S. Ameh, S. C. Ikpeseni, and L. S. Lawal, “A Review of Field Corrosion Control and Monitoring Techniques of the Upstream Oil and Gas Pipelines,” *Niger. J. Technol. Dev.*, 2018, doi: 10.4314/njtd.v14i2.5.

[53] A. Peratta, J. Baynham, R. Adey, and G. F. Pimenta, “Intelligent remote monitoring system for cathodic protection of transmission pipelines,” in *NACE - International Corrosion Conference Series*, 2009.

[54] F. J. Hoppe, S. E. Turner, S. P. Basu, and G. E. Rogers, “Design, installation and field experience with real-time cathodic protection monitoring of pipe-type cable systems,” in *Proceedings of the IEEE Power Engineering Society Transmission and Distribution Conference*, 1996, doi: 10.1109/tdc.1996.547581.

[55] F. Abate, D. Di Caro, G. Di Leo, and A. Pietrosanto, “A Networked Control System for Gas Pipeline Cathodic Protection,” *IEEE Trans. Instrum. Meas.*, 2020, doi: 10.1109/TIM.2019.2906968.

[56] A. Kara, M. A. Al Imran, and K. Karadag, “Linear Wireless Sensor Networks for Cathodic Protection Monitoring of Pipelines,” in *Proceedings of the 2019 International Conference on Mechatronics, Robotics and Systems Engineering, MoRSE 2019*, 2019, doi: 10.1109/MoRSE48060.2019.8998664.

[57] NACE International, *NACE Pipeline Corrosion Integrity Management - Training Manual*. Houston, 2011.

[58] American Society of Mechanical Engineers, “ASME B31.8: Gas Transmission and Distribution Piping Systems,” *Am. Soc. Mech. Eng.*, 2003, doi: 10.1520/G0154-12A.

[59] CSA Group, *Oil and gas pipeline systems. CSA Z662*. 2015.

[60] J. Woodhouse, R. Davies, M. Mustafa, U. Bryan, and P. Jay, “Pas 55-1:2008,” *Br. Stand.*, 2008.

[61] NACE International, “International Measure of Prevention, Application and Economics of Corrosion Technologies Study,” 2016. [Online]. Available: http://impact.nace.org/documents/Nace-International-Report.pdf. [Accessed: 29-Aug-2020].

[62] ISO, *ISO 13372:2012(en), Condition monitoring and diagnostics of machines — Vocabulary*. 2012.

[63] P. D. T. O’Connor and A. Kleyner, *Practical Reliability Engineering: Fifth Edition*.

2011.

[64] R. Ahmad and S. Kamaruddin, “An overview of time-based and condition-based maintenance in industrial application,” *Comput. Ind. Eng.*, 2012, doi: 10.1016/j.cie.2012.02.002.

[65] K. F. Martin, “A review by discussion of condition monitoring and fault diagnosis in machine tools,” *Int. J. Mach. Tools Manuf.*, 1994, doi: 10.1016/0890-6955(94)90083-3.

[66] ISO, “ISO 17359:2018,” 2006.

[67] ISO, “ISO 13374-2:2007(en) Condition monitoring and diagnostics of machines — Data processing, communication and presentation — Part 2: Data processing,” 2002. [Online]. Available: https://www.iso.org/obp/ui/#iso:std:iso:13374:-2:ed-1:v2:en. [Accessed: 30-Sep-2020].

[68] K. Medjaher and N. Zerhouni, “Framework for a hybrid prognostics,” in *Chemical Engineering Transactions*, 2013, doi: 10.3303/CET1333016.

[69] A. Ragab, M. S. Ouali, S. Yacout, and H. Osman, “Remaining useful life prediction using prognostic methodology based on logical analysis of data and Kaplan–Meier estimation,” *J. Intell. Manuf.*, 2016, doi: 10.1007/s10845-014-0926-3.

[70] P. Gackowiec, “General overview of maintenance strategies – concepts and approaches,” *Multidiscip. Asp. Prod. Eng.*, 2019, doi: 10.2478/mape-2019-0013.

[71] L. Hitchcock, “ISO standards for condition monitoring,” in *Proceedings of the 1st World Congress on Engineering Asset Management, WCEAM 2006*, 2006, doi: 10.1007/978-1-84628-814-2_65.

[72] N. Sakib and T. Wuest, “Challenges and Opportunities of Condition-based Predictive Maintenance: A Review,” in *Procedia CIRP*, 2018, doi: 10.1016/j.procir.2018.08.318.

[73] T. Xu, T. Tang, H. Wang, and T. Yuan, “Risk-based predictive maintenance for safety-critical systems by using probabilistic inference,” *Math. Probl. Eng.*, 2013, doi: 10.1155/2013/947104.

[74] E. Arzaghi, M. M. Abaei, R. Abbassi, V. Garaniya, C. Chin, and F. Khan, “Risk-based maintenance planning of subsea pipelines through fatigue crack growth monitoring,” *Eng. Fail. Anal.*, 2017, doi: 10.1016/j.engfailanal.2017.06.003.

[75] M. Rausand, “Reliability centered maintenance,” in *Marine Technology and Engineering*, 2011, doi: 10.2307/1268924.

[76] R. G. Jansen, L. F. Wiertz, E. S. Meyer, and L. P. J. J. Noldus, “Reliability analysis of observational data: Problems, solutions, and software implementation,” in *Behavior Research Methods, Instruments, and Computers*, 2003, doi: 10.3758/BF03195516.

[77] K. Ramasubramanian and A. Singh, *Machine learning using R*. 2016.

[78] A. Dey, “Machine Learning Algorithms: A Review,” *Int. J. Comput. Sci. Inf.*

Technol., 2016.

[79] A. Viti, A. Terzi, and L. Bertolaccini, “A practical overview on probability distributions,” *J. Thorac. Dis.*, 2015, doi: 10.3978/j.issn.2072-1439.2015.01.37.

[80] R. D. Shachter and M. Alan Peot, “Decision Making Using Probabilistic Inference Methods,” in *Proceedings of the Eighth Conference on Uncertainty in Artificial Intelligence*, 2013.

[81] M. Kuhn and K. Johnson, *Applied predictive modeling*. 2013.

[82] M. Paolanti, L. Romeo, A. Felicetti, A. Mancini, E. Frontoni, and J. Loncarski, “Machine Learning approach for Predictive Maintenance in Industry 4.0,” in *2018 14th IEEE/ASME International Conference on Mechatronic and Embedded Systems and Applications, MESA 2018*, 2018, doi: 10.1109/MESA.2018.8449150.

[83] M. Berry, A. Mohamed, and B. W. Yap, “Supervised and Unsupervised Learning for Data Science,” *Unsupervised Semi-Supervised Learn.*, 2019, doi: 10.1007/978-3-030-22475-2.

[84] V. A. Barbur, D. C. Montgomery, and E. A. Peck, “Introduction to Linear Regression Analysis.,” *Stat.*, 1994, doi: 10.2307/2348362.

[85] T. G. Clark, M. J. Bradburn, S. B. Love, and D. G. Altman, “Survival Analysis Part I: Basic concepts and first analyses,” *Br. J. Cancer*, 2003.

[86] G. A. Susto, A. Schirru, S. Pampuri, S. McLoone, and A. Beghi, “Machine learning for predictive maintenance: A multiple classifier approach,” *IEEE Trans. Ind. Informatics*, 2015, doi: 10.1109/TII.2014.2349359.

[87] H. A. Gohel, H. Upadhyay, L. Lagos, K. Cooper, and A. Sanzetenea, “Predictive maintenance architecture development for nuclear infrastructure using machine learning,” *Nucl. Eng. Technol.*, 2020, doi: 10.1016/j.net.2019.12.029.

[88] P. F. Orrù, A. Zoccheddu, L. Sassu, C. Mattia, R. Cozza, and S. Arena, “Machine learning approach using MLP and SVM algorithms for the fault prediction of a centrifugal pump in the oil and gas industry,” *Sustain.*, 2020, doi: 10.3390/su12114776.

[89] O. D. Apuke, “Quantitative Research Methods : A Synopsis Approach,” *Kuwait Chapter Arab. J. Bus. Manag. Rev.*, 2017, doi: 10.12816/0040336.

[90] P. D. Leedy and J. E. Ormrod, *Practical Research: Planning and design (11th ed.)*. 2010.

[91] L. Given, *The SAGE Encyclopedia of Qualitative Research Methods*. 2455 Teller Road, Thousand Oaks California 91320 United States: SAGE Publications, Inc., 2008.

[92] R. Wieringa, “Empirical research methods for technology validation: Scaling up to practice,” *J. Syst. Softw.*, 2014, doi: 10.1016/j.jss.2013.11.1097.

[93] A. J. Hugo, “Estimation of alarm deadbands,” in *IFAC Proceedings Volumes (IFAC-PapersOnline)*, 2009, doi: 10.3182/20090630-4-ES-2003.0054.

[94] S. S. B, K. P. Pranav, and K. Raj R., “Solar powered corrosion prevention in iron pipelines using Impressed Current Cathodic Protection,” in *2014 14th International Conference on Environment and Electrical Engineering, EEEIC 2014 - Conference Proceedings*, 2014, doi: 10.1109/EEEIC.2014.6835896.

[95] W. L. Neuman, *Social research methods: qualitative and quantitative approaches: Boston, [Mass.]: Pearson*. 2011.

[96] K. Williamson and G. Johanson, *Research Methods: Information, Systems, and Contexts: Second Edition*. 2017.

[97] O. D. Anderson, “More effective time-series analysis and forecasting,” *J. Comput. Appl. Math.*, 1995, doi: 10.1016/0377-0427(95)00011-9.

[98] I. Madanhire and C. Mbohwa, “Application of Statistical Process Control (SPC) in Manufacturing Industry in a Developing Country,” in *Procedia CIRP*, 2016, doi: 10.1016/j.procir.2016.01.137.

[99] A. Saxena and A. Saad, “Evolving an artificial neural network classifier for condition monitoring of rotating mechanical systems,” *Appl. Soft Comput. J.*, 2007, doi: 10.1016/j.asoc.2005.10.001.

[100] ISO13381-1, “Condition Monitoring and Diagnostics of Machines —Prognostics —Part 1: General Guidelines,” *International Standard, ISO*. 2015, doi: 10.1109/TDEI.2015.7076829.

[101] T. Chai and R. R. Draxler, “Root mean square error (RMSE) or mean absolute error (MAE)? -Arguments against avoiding RMSE in the literature,” *Geosci. Model Dev.*, 2014, doi: 10.5194/gmd-7-1247-2014.

[102] C. C. Terry M Therneau,Thomas Lumley,Atkinson Elizabeth, “Survival Analysis,” *CRAN*, 2020. [Online]. Available: https://cran.r-project.org/web/packages/survival/survival.pdf. [Accessed: 03-Oct-2020].

[103] G. Thomas, “Methodology part 2:the design frame,” in *How to do your research Project*, 2013.

[104] D. Theofanidis and A. Fountouki, “Limitations and Delimitations in the Research Process,” *Perioper. Nurs.*, 2018, doi: 10.5281/zenodo.2552022.

[105] J. Brittain, M. Cendon, J. Nizzi, and J. Pleis, “Data Scientist’s Analysis Toolbox: Comparison of Python, R, and SAS Performance,” *SMU Data Sci. Rev.*, 2018.

[106] R. H. Shumway and D. S. Stoffer, *Time series analysis and its applications With R Examples EZ Edition*. 2016.

[107] Rob Hyndman *et al.*, “Forecasting Functions for Time Series and Linear Models,” *CRAN*, 2020. [Online]. Available: https://cran.r-project.org/web/packages/forecast/forecast.pdf. [Accessed: 07-Oct-2020].

[108] W. K. Härdle and L. Simar, *Applied multivariate statistical analysis*. 2013.

[109] C. M. Krebsbach, “EQS Output Conversion to lavaan Functions,” *CRAN*, 2013. [Online]. Available: https://cran.r-project.org/web/packages/eqs2lavaan/eqs2lavaan.pdf. [Accessed: 10-Oct-2020].

[110] M. H. Saadat, “Steady state analysis of power systems including the effects of control devices,” *Electr. Power Syst. Res.*, 1979, doi: 10.1016/0378-7796(79)90016-6.

[111] W. Krämer, “Durbin–Watson Test,” in *International Encyclopedia of Statistical Science*, Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 408–409.

[112] T. Thadewald and H. Büning, “Jarque-Bera test and its competitors for testing normality - A power comparison,” *J. Appl. Stat.*, 2007, doi: 10.1080/02664760600994539.

[113] M. Kuhn, “The caret Package,” *http://topepo.github.io/caret/index.html*, 2019. [Online]. Available: http://topepo.github.io/caret/models-clustered-by-tag-similarity.html. [Accessed: 25-Oct-2020].

[114] A. Gelman and J. Hill, *Data Analysis Using Regression and Multilevel/Hierarchical Models*. 2006.

[115] Robert Baboian, *NACE Corrosion Engineer’s Reference Book*. 1980.

[116] H. Wang, C. Du, Z. Liu, L. Wang, and D. Ding, “Effect of alternating current on the cathodic protection and interface structure of X80 steel,” *Materials (Basel).*, 2017, doi: 10.3390/ma10080851.

[117] M. Chen, S. Liu, J. Zhu, C. Xie, H. Tian, and J. Li, “Effects and characteristics of AC interference on parallel underground pipelines caused by an AC electrified railway,” *Energies*, 2018, doi: 10.3390/en11092255.

APPENDIX A

A1: Standard EMF Series

Standard EMF Series		
	Metal	**Standard Electrode Potential V°**
Increasingly Cathodic >>	**Au**	+1.420V
	Cu	+ 0.34 V
	Pb	-0.126 V
	Sn	-0.136 V
	Ni	-0.25 V
	Co	-0.277 V
	Cd	-0.403 V
<< Increasingly Anodic	**Fe**	-0.44 V
	Cr	-0.744 V
	Zn	-0.763 V
	Al	-1.662 V
	Mg	-2.262 V
	Na	-2.714 V
	K	-2.924 V

Table A0-1 - Standard EMF Series - Source: Adapted from [27]

A2: Galvanic Series

Galvanic Series	
	Metal
Increasingly Cathodic >>	Platinum
	Gold
	Graphite
	Titanium
	Silver
	316 Stainless steel (passive)
	304 Stainless steel (passive)
	Inconel (80Ni–13Cr–7Fe) (passive)
	Nickel (passive)
	Monel (70Ni–30Cu)
	Copper-nickel alloys
<< Increasingly Anodic	Bronzes (Cu–Sn alloys)
	Copper
	Brasses (Cu–Zn alloys)
	Inconel (active)
	Nickel (active)
	Tin
	Lead
	316 Stainless steel (active)
	304 Stainless steel (active)

Galvanic Series	
	Metal
	Cast iron
	Iron and steel
	Aluminium alloys
	Cadmium
	Commercially pure aluminium
	Zinc
	Magnesium and magnesium alloys

Table A0-2 - Galvanic Series - Source: Adapted from [22]

A3: Relative Potentials of Common RE Against SHE

Relative Potentials of Common Reference Electrodes vs the Saturated Hydrogen Electrode	
Electrode (Half-Cell)*	**Potential (V)**
Standard Hydrogen	0.000V
Copper-Copper Sulfate (CSE)	+0.32V
Saturated Silver-Silver Chloride (SSC)	+0.23V
Saturated Calomel (SCE)	+0.23V
Zinc (ZRE)	–0.78V

Table A0-3 - Relative DC Potentials of RE vs SHE - Source: Adapted from [31]

A4: Convert RE Potentials

Park suggests the following formula for adjusting readings based on the RE used (this is for a CSE) [34]:

$$P_A = P_{RE} - P_{CSE} + P_{Reading\ vs\ RE} \quad \textit{0.1 Adjusted DC Potential for CSE}$$

Where:

- P_A = Adjusted reading for CSE in V_{DC}
- P_{RE} = Electrode potential of RE in use in V_{DC}
- P_{CSE} = Electrode potential of CSE in V_{DC}
- $P_{Reading\ vs\ RE}$ = Voltage taken by RE in use in V_{DC}

An experiment conducted by Park illustrates the conversion of metal DC potential measurements when using different RE's. Defining a native potential of iron as -440mVDC, the table below shows each RE's DC potential, the expected metal DC voltage reading when using each RE, and the adjusted expected readings (V_{CSE}) for each RE [29].

Reference Electrode Types and ΔP for Fe					
Electrode Composition	**Chemical Formula**	**Environment**	**Voltage Difference from SHE @ 25 Degrees Celsius**	**ΔP (Expected Voltage Reading) for Native Fe Potential**	**Expected Readings Adjusted to the CSE Potential**
Standard Hydrogen Electrode (SHE)	SHE	Primary Reference Electrode	0mV	-440mV	N/A
Copper-Copper Sulfate Electrode (CSE)	$Cu/CuSO_4$	Underground Structures	+316mV	-440mV - 316mV = -756mV	N/A
Silver-Silver Chloride Reference Electrode (SSC)	Ag/AgCl	Seawater and Concrete Structures	+199mV	-440mV - 199mV = -639 V	199mV - 316mV + reading = [199 - 316 + (-639)]mV = -756mV
Calomel Reference Electrode (SCE)	Calomel	Laboratory	+244mV	-440mV - 244mV = -684mV	244mV - 316mV + reading = [244 - 316 + (-684)]mV = -756mV
Zinc Reference Electrode (ZRE)	Zn	Pseudo-Reference (Underground Structures)	-762mV	-440mV - (-762mV) = 322mV	-762mV - 316mV + reading = [-762 - 316 + 322]mV = -756mV

Table A0-4 - RE Types and ΔP for Fe - Source: Adapted from [29]

A5: Corrosion Rate Methods

This section evaluates either the weight-loss rate per surface area over time, the corrosion penetration rate or the electrochemical rate methods [32].

Electrical Resistance Method

Corrosion of a metal results in a change of the electrical resistance over time as the metal size reduce. The expected electrical resistance is inversely proportional to the metal size (resistance will increase if metal size decrease). Typical instruments monitor corrosion progression rather than the corrosion rate [32].

The formula below describes the electrical resistance in terms of its resistivity, wire length and surface area of the metal [25]:

$$\mathbf{R} = \frac{\boldsymbol{\rho\ell}}{\mathbf{A}} \qquad \textit{0.2 – Electrical Resistivity (Ohm)}$$

Where:

- R = Resistivity in Ω
- ρ = Resistivity of specific metal
- l = Wire length
- A = Area of metal

Mass Loss Rate

According to Baboian, the corrosion rate is directly proportional to the current flow and current density. Faraday's law can be used in this calculation to determine the mass-loss rate based on current magnitude [115]:

$$\boldsymbol{MR = K_2 \times I_{Corr} \times EW} \qquad \textit{0.3 – Mass Loss Rate (g/m}^2\textit{d)}$$

Where:

- MR = Mass Loss Rate
- EW = Equivalent Weight
- K_2 = Constant that defines the units of the corrosion rate
- I_{Corr} = Current Density in A/m^2 or µA/cm^2

Corrosion Rate Expressed Using Current Density

Callister further defines the corrosion rate in terms of the surface area, the number of electrons related to the ionization process of each metal atom and Faraday's constant [22].

$$\mathbf{r} = \frac{\boldsymbol{i}}{\boldsymbol{n\mathfrak{F}}} \qquad \textit{0.4 – Corrosion Rate (mol/m}^2\textit{-s)}$$

Where:

- r = Corrosion Rate
- i = Current per unit surface area of corroding material
- n = Number of electrons
- F = 96,500 C/mol

Corrosion Penetration Rate

The corrosion penetration rate can be calculated using the following equation [22]:

$$\boldsymbol{CPR = \frac{KW}{pAt}} \qquad \textit{0.5 – Corrosion Penetration Rate(mm/yr)}$$

Where:

- W = Weight loss after exposure time t
- K = Constant that defines the units of the corrosion rate
- ρ = Density in grams/cm³
- A = Area of the sample in cm²
- t = Unit of Time

In summary, the formula used for corrosion rate depends on the data available and the expression of corrosion rate for specific variables [22].

A6: CP System Characteristics

Metwally [49] summarized the characteristics of ICCP and galvanic CP systems as follows:

CP System Characteristics		
Characteristic	**Galvanic CP**	**ICCP**
External power	No	Yes
Driving voltage	Fixed	Variable
Current required	Limited and Low	Variable and High
Soil conductivity	High	Wide Range
Interference	Negligible	Possible and Risky

Table A0-5 - CP System Characteristics - Source: Adapted from [49]

A7: Stray Current Density Calculation

To calculate the current density of DC interference, Metwally et al. provide the following formula [48][49]:

$$\mathbf{i} = \boldsymbol{\sigma_e e} \qquad \textit{0.6 – Current Density (mA/m}^2\textit{)}$$

Where:

- i = Current Density in mA/m^2
- σ_e = Electrical conductivity of soil in S/m
- e = Electric field in V/m

NACE suggests that instantaneous potentials at a single location along the pipeline is not a true reflection of the magnitude of the stray current, but can be used for reference to determine if any stray current is present [46]. Holtsbaum also suggests that a 20% potential shift from steady-state potentials requires investigation [44].

Wang suggests that AC interference can exist where pipelines run parallel or cross AC railways or high-voltage AC transmission lines [116]. Electromagnetic coupling, instantaneous or steady-state, with the pipeline or structure, can cause AC stray current corrosion. Dawalibi provides methods for calculating inductive coupling of steady-state AC currents, while Bortles and Christoforidis developed finite methods for calculating the inductive interference under normal and fault conditions [117].

NACE provides the following formula for calculating the AC density for AC interference [46]:

$$\mathbf{i_{ac}} = \frac{(\mathbf{8}\boldsymbol{V_{ac}})}{\boldsymbol{\rho\pi d}} \qquad \textit{0.7 – AC Current Density (A}_{AC}\textit{/m}^2\textit{)}$$

Where:

- i_{ac} = AC current density (A_{AC}/m^2)
- V_{ac} = AC voltage-to-earth (V_{AC})
- ρ = Soil resistivity (Ω-m)
- d = Coating holiday diameter (m)

NACE further suggests that grounding of pipelines if the AC density exceeds 30 A/m^2 and regulatory grounding if the AC density exceeds 100 A/m^2 [46].

A8: CP Inspection Methods

ILI PURPOSE	METAL-LOSS TOOLS			CRACK-DETECTION TOOLS		CALIPER TOOLS	MAPPING TOOLS
	Magnetic Flux Leakage (MFL)		Ultrasonic (compression wave)	Ultrasonic (shear wave)	Transverse MFL		
	Standard-resolution (SR) MFL	High-resolution (HR) MFL					
METAL LOSS (CORROSION) External corrosion Internal corrosion	detection,[A] sizing,[B] no ID/OD[C] discrimination	detection,[A] sizing[B]	detection,[A] sizing[B]	detection,[A] sizing[B]	detection,[A] sizing[B]	no detection	no detection
NARROW AXIAL EXTERNAL CORROSION	no detection[A]	no detection[A]	detection,[A] sizing[B]	detection,[A] sizing[B]	detection,[A] sizing[B]	no detection	no detection
CRACKS AND CRACK-LIKE DEFECTS (Axial) Stress corrosion cracking Fatigue cracks Longitudinal seam weld imperfections Incomplete fusion (lack of fusion) Toe cracks	no detection	no detection	no detection	detection,[A] sizing[B]	detection,[A][D] sizing[B]	no detection	no detection
CIRCUMFERENTIAL CRACKING	no detection	detection,[D] sizing[D]	no detection	detection,[A] sizing[B] if modified[E]	no detection	no detection	no detection
DENTS SHARP DENTS WRINKLE BENDS BUCKLES	detection[F]	detection,[F] sizing not reliable	detection,[F] sizing not reliable	detection,[F] sizing not reliable	detection,[F] sizing not reliable	detection,[G] sizing	detection, sizing not reliable
	In case of detection, circumferential position is provided.						
GOUGES	Detection[H] and Sizing[H]						no detection
LAMINATION OR INCLUSION	limited detection	limited detection	detection, sizing[B]	detection, sizing[B]	limited detection	no detection	no detection
PREVIOUS REPAIRS	detection of steel sleeves and patches, others only with ferrous markers		detection only of steel sleeves and patches welded to pipe	detection only of steel sleeves and patches welded to pipe	detection only of steel sleeves and patches, others only with ferrous markers	no detection	no detection
MILL-RELATED ANOMALIES	limited detection	limited detection	detection	detection	limited detection	no detection	no detection
BENDS	no detection	no detection	no detection	no detection	no detection	detection, sizing[H]	detection, sizing
OVALITIES	no detection	no detection	no detection	no detection	no detection	detection, sizing[E]	detection, sizing[E][I]
PIPELINE COORDINATES	no detection	no detection	no detection	no detection	no detection	no detection	detection, sizing

(A) Limited by the minimum detectable depth, length, and width of the defects.
(B) Defined by the specified sizing accuracy of the tool.
(C) Internal diameter (ID) and outside diameter (OD).
(D) Reduced probability of detection (POD) for tight cracks.
(E) Transducers to be rotated by 90°.
(F) Reduced reliability depending on the size and shape of the dent.
(G) Depending on the configuration of the tool, also circumferential position.
(H) If equipped for bend measurements.
(I) If the tool is equipped for ovality measurement.

Shaded area indicates ILI technologies that can be used only in liquid environments, i.e., liquids pipelines or in gas pipelines with a liquid couplant.

Table A0-6 - NACE Inspection Methods - Source: Adapted from [57]

A9: Rectifier Maintenance Flowchart

Holtsbaum provides the following flow diagram for identifying ICCP faults:

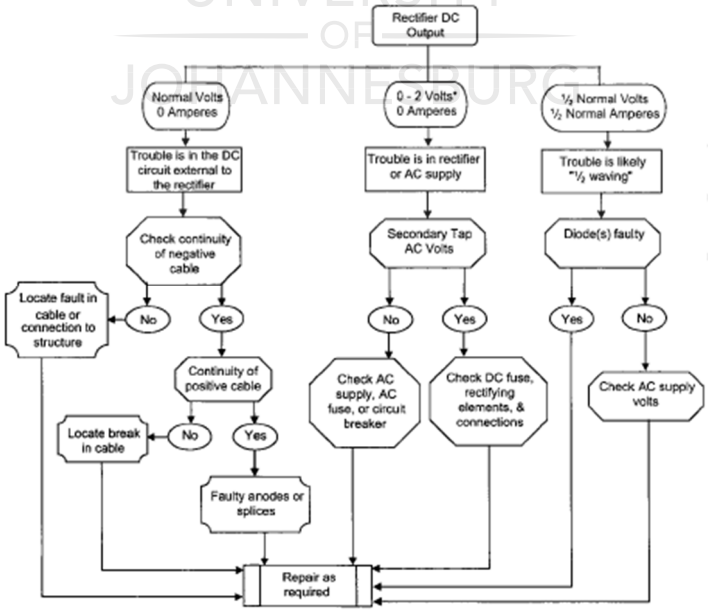

Figure A0-1 - Flow Diagram for Rectifier Faults - Source: Adapted from[44]

APPENDIX B

B1: CP Health Indicator

TP Health Indicator Calculation

The table below indicates the various conditions for the TP health indication:

TP Health Indicator Calculation														
ID	Unit Type	Unit Risk Factor	Rectifier On	Risk Level OP	Risk Level UP	Risk Level P	OP Factor	UP Factor	P Factor	Distance Factor	Total Distance	Risk Value	Total Possible Score	Health Value
1	1	1	1	0	0	0	2	3	1.15	1	21	0.00	15	100.00%
2	1	1	1	1	0	0	2	3	1.15	1	21	2.10	15	86.03%
3	1	1	1	2	0	0	2	3	1.15	1	21	4.19	15	72.06%
4	1	1	1	3	0	0	2	3	1.15	1	21	6.29	15	58.10%
5	1	1	1	4	0	0	2	3	1.15	1	21	8.38	15	44.13%
6	1	1	0	0	0	0	2	3	1.15	1	21	2.10	15	86.03%
7	1	1	0	1	0	0	2	3	1.15	1	21	4.19	15	72.06%
8	1	1	0	2	0	0	2	3	1.15	1	21	6.29	15	58.10%
9	1	1	0	3	0	0	2	3	1.15	1	21	8.38	15	44.13%
10	1	1	0	4	0	0	2	3	1.15	1	21	10.48	15	30.16%
11	1	1	1	0	0	0	2	3	1.15	1	21	0.00	15	100.00%
12	1	1	1	0	1	0	2	3	1.15	1	21	3.14	15	79.05%
13	1	1	1	0	2	0	2	3	1.15	1	21	6.29	15	58.10%
14	1	1	1	0	3	0	2	3	1.15	1	21	9.43	15	37.14%
15	1	1	1	0	4	0	2	3	1.15	1	21	12.57	15	16.19%
16	1	1	0	0	0	0	2	3	1.15	1	21	2.10	15	86.03%
17	1	1	0	0	1	0	2	3	1.15	1	21	5.24	15	65.08%
18	1	1	0	0	2	0	2	3	1.15	1	21	8.38	15	44.13%
19	1	1	0	0	3	0	2	3	1.15	1	21	11.52	15	23.17%
20	1	1	0	0	4	0	2	3	1.15	1	21	14.67	15	2.22%
21	1	1	1	0	0	1	2	3	1.15	1	21	1.20	15	91.97%
22	1	1	0	0	0	1	2	3	1.15	1	21	3.30	15	78.00%

Table B0-1 - TP Health Indicator

TRU Health Indicator Calculation

The table below indicates the various conditions for the TRU health indication:

TRU Health Indicator Calculation														
ID	Unit Type	Unit Risk Factor	Rectifier On	Risk Level OP	Risk Level UP	Risk Level P	OP Factor	UP Factor	P Factor	Distance Factor	Total Distance	Risk Value	Total Possible Score	Health Value
1	2	2	1	0	0	0	2	3	1.15	1	21	0.00	17	100.00%
2	2	2	1	1	0	0	2	3	1.15	1	21	2.10	17	87.68%
3	2	2	1	2	0	0	2	3	1.15	1	21	4.19	17	75.35%
4	2	2	1	3	0	0	2	3	1.15	1	21	6.29	17	63.03%
5	2	2	1	4	0	0	2	3	1.15	1	21	8.38	17	50.70%
6	2	2	0	0	0	0	2	3	1.15	1	21	4.19	17	75.35%
7	2	2	0	1	0	0	2	3	1.15	1	21	6.29	17	63.03%
8	2	2	0	2	0	0	2	3	1.15	1	21	8.38	17	50.70%
9	2	2	0	3	0	0	2	3	1.15	1	21	10.48	17	38.38%
10	2	2	0	4	0	0	2	3	1.15	1	21	12.57	17	26.05%
11	2	2	1	0	0	0	2	3	1.15	1	21	0.00	17	100.00%
12	2	2	1	0	1	0	2	3	1.15	1	21	3.14	17	81.51%
13	2	2	1	0	2	0	2	3	1.15	1	21	6.29	17	63.03%
14	2	2	1	0	3	0	2	3	1.15	1	21	9.43	17	44.54%
15	2	2	1	0	4	0	2	3	1.15	1	21	12.57	17	26.05%
16	2	2	0	0	0	0	2	3	1.15	1	21	4.19	17	75.35%
17	2	2	0	0	1	0	2	3	1.15	1	21	7.33	17	56.86%
18	2	2	0	0	2	0	2	3	1.15	1	21	10.48	17	38.38%
19	2	2	0	0	3	0	2	3	1.15	1	21	13.62	17	19.89%
20	2	2	0	0	4	0	2	3	1.15	1	21	16.76	17	1.40%
21	2	2	1	0	0	1	2	3	1.15	1	21	1.20	17	92.91%
22	2	2	0	0	0	1	2	3	1.15	1	21	5.40	17	68.26%

Table B0-2 - TRU Health Indication

FDU Health Indicator Calculation

The table below indicates the various conditions for the FDU health indication:

FDU Health Indicator Calculation														
ID	Unit Type	Unit Risk Factor	Rectifier On	Risk Level OP	Risk Level UP	Risk Level P	OP Factor	UP Factor	P Factor	Distance Factor	Total Distance	Risk Value	Total Possible Score	Health Value
1	3	3	1	0	0	0	2	3	1.15	1	21	0.00	19	100.00%
2	3	3	1	1	0	0	2	3	1.15	1	21	2.10	19	88.97%
3	3	3	1	2	0	0	2	3	1.15	1	21	4.19	19	77.94%
4	3	3	1	3	0	0	2	3	1.15	1	21	6.29	19	66.92%
5	3	3	1	4	0	0	2	3	1.15	1	21	8.38	19	55.89%
6	3	3	0	0	0	0	2	3	1.15	1	21	6.29	19	66.92%
7	3	3	0	1	0	0	2	3	1.15	1	21	8.38	19	55.89%
8	3	3	0	2	0	0	2	3	1.15	1	21	10.48	19	44.86%
9	3	3	0	3	0	0	2	3	1.15	1	21	12.57	19	33.83%
10	3	3	0	4	0	0	2	3	1.15	1	21	14.67	19	22.81%
11	3	3	1	0	0	0	2	3	1.15	1	21	0.00	19	100.00%
12	3	3	1	0	1	0	2	3	1.15	1	21	3.14	19	83.46%
13	3	3	1	0	2	0	2	3	1.15	1	21	6.29	19	66.92%
14	3	3	1	0	3	0	2	3	1.15	1	21	9.43	19	50.38%
15	3	3	1	0	4	0	2	3	1.15	1	21	12.57	19	33.83%
16	3	3	0	0	0	0	2	3	1.15	1	21	6.29	19	66.92%
17	3	3	0	0	1	0	2	3	1.15	1	21	9.43	19	50.38%
18	3	3	0	0	2	0	2	3	1.15	1	21	12.57	19	33.83%
19	3	3	0	0	3	0	2	3	1.15	1	21	15.71	19	17.29%
20	3	3	0	0	4	0	2	3	1.15	1	21	18.86	19	0.75%
21	3	3	1	0	0	1	2	3	1.15	1	21	1.20	19	93.66%
22	3	3	0	0	0	1	2	3	1.15	1	21	7.49	19	60.58%

Table B0-3 - FDU Health Indication

The graph below indicates the health indicators for the different equipment defined (TRU, FDU and TP). As shown on the graph, a clear distinction is made between unit types and risk levels.

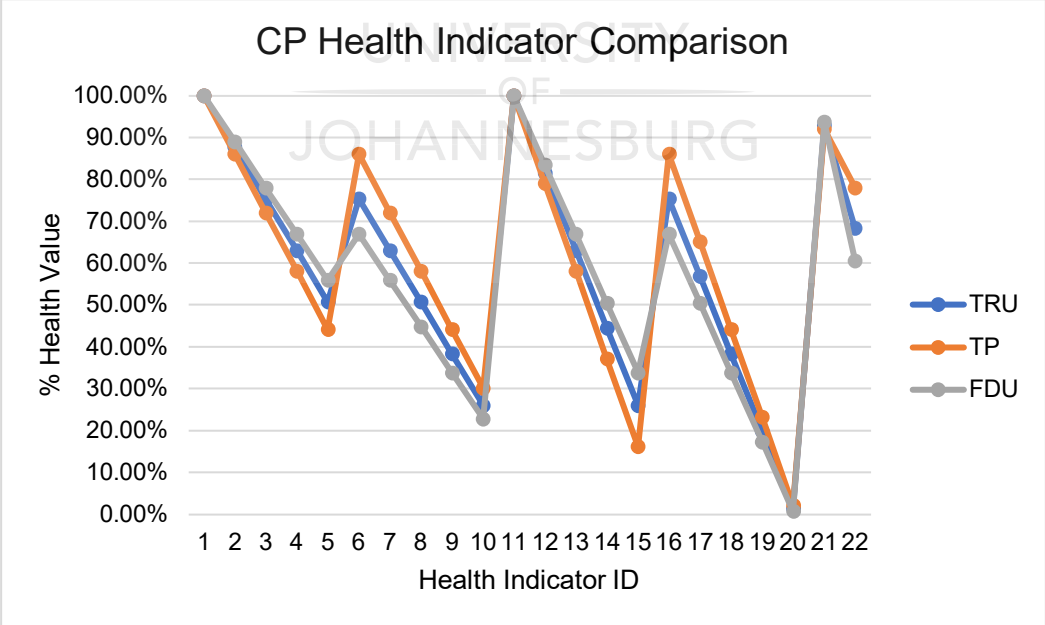

Figure B0-1 - CP Health Indicator Comparison

A visual representation of a 9-km pipeline section is illustrated below with the individual unit health indicator and the overall pipeline health indicator for a state of under-protection, with risk level 2, at the TRU. Percentage values are deduced from the tables above by substituting the TP location.

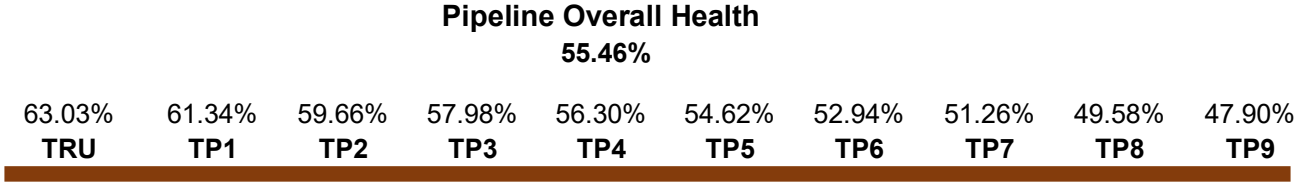

Figure B0-2 - CP Health Indicator Visual Representation

APPENDIX C

C1: Data Exploration Methods

Statistical Methods Evaluated							
Chapter	**Method**	**Aim**	**ML Model**	**Performance Metric**	**Performance Result**	**Research Objective**	**Research Question**
Data Exploration	Plotting ICCP PV's on a line graph	Visual review of the raw data.	None	None	N/A	N/A	1
Data Exploration	Plotting CP pipe potential with OW	Visual evaluation of the CP pipe potential with reference to the defined OW.	None	None	N/A	N/A	1
Data Exploration	Plotting CP pipe potential with spikes and OW	Visual evaluation of the CP pipe potential for high magnitude spikes (up or down).	None	None	N/A	N/A	1
Data Exploration	Plotting PV Distribution Densities	Visually inspect the distribution type, skewness and kurtosis of data set.	None	None	N/A	N/A	1,2
Data Exploration	Correlation Diagrams	Determine the correlation between variables.	None	None	N/A	N/A	1
Data Exploration	Time-series Decomposition of CP pipe potential	Determine the CP pipe potential trend with noise removed for one or more instances.	None	None	N/A	N/A	1
Data Exploration	Long-term Time Series Analysis	Evaluated the difference in hourly, daily, weekly, monthly and quarterly CP pipe potential trends to inform maintenance required.	None	None	N/A	N/A	1,2
Data Exploration	Long-term Time Series Analysis - Forecasting	Forecast CP pipe potentials on a TS trend component to determine the estimated potential per TS window.	None	None	N/A	N/A	1,2

Table C0-1 - Data Analysis Techniques for Data Exploration

C2: Statistical Methods For Predictive Modelling

Statistical Methods Evaluated							
Chapter	**Method**	**Aim**	**ML Model**	**Performance Metric**	**Performance Result**	**Research Objective**	**Research Question**
Data Exploration	CP Health Indicator Calculation	Calculate and determine the CP health status per unit based on data received and estimation	None	% Error	3%	1	1,2
Data Exploration	Descriptive Statistics	Describe conformance to OW as % or time statistics	None	None	N/A	1	1,2
Predictive Modelling	TRU CP pipe potential Prediction at Steady-State	Predict CP pipe potential of CP TRU at steady-state	RF	RMSE	0.153	1	1
Predictive Modelling	Malfunctioning FDU CP pipe potential Prediction	Predict CP pipe potential of a malfunctioning CP FDU	RF	RMSE	26.95	1	1
Predictive Modelling	FDU CP pipe potential Prediction with Stray Current	Predict CP pipe potential of CP FDU with Stray Current	RF	RMSE	1.926	1	1
Predictive Modelling	FDU CP pipe potential Prediction with Increased Sampling Interval	Determine if Prediction Error is improved by increasing the sampling interval	RF	RMSE	0.675	1	1
Predictive Modelling	Estimate TP potentials using LR	Estimate downstream TP potentials by estimating output current coefficients for TP's and using LR to estimate potentials.	RF	% Accuracy	99.95%	1	1
Predictive Modelling	CP pipe potential State Prediction at Steady-State	Predict the CP pipe potential State using Classification.	RF	% Accuracy	98.87%	1	1
Predictive Modelling	CP pipe potential State Prediction with Stray Current	Predict the CP pipe potential State with Stray Current using Classification.	RF	% Accuracy	93.66%	1	1

Table C0-2 - Statistical Methods State Prediction

C3: Statistical Methods For Maintenance Suggestion

Statistical Methods Evaluated							
Chapter	**Method**	**Aim**	**ML Model**	**Performance Metric**	**Performance Result**	**Research Objective**	**Research Question**
Data Exploration	Long-term Time Series Analysis	Evaluated the difference in hourly, daily, weekly, monthly and quarterly CP pipe potential trends to inform maintenance required.	None	None	N/A	2	1,2
Data Exploration	Long-term Time Series Analysis - Forecasting	Forecast CP pipe potentials on a TS trend component to determine the estimated potential per TS window.	None	None	N/A	2	1,2
Predictive Modelling	Time-to-State Prediction using Survival Package	Determine state survival times using KM-curve.	None	None	N/A	2	1
Predictive Modelling	Time-to-State Prediction using Cycle Times	Determine the Time-to-Event using predefined cycle times per state.	None	None	N/A	2	1
Predictive Modelling	Maintenance Suggestion at Steady-State	Predict maintenance activities based on risk levels.	svmLinear	% Accuracy	99.67%	2	1
Predictive Modelling	Maintenance Suggestion with Stray Current	Predict maintenance activities based on risk levels.	RF	% Accuracy	96.64%	2	1

Table C0-3 - Statistical Methods for Maintenance Suggestion

APPENDIX D

D1: List Of R Packages Used In This Study

- data.table
- readr
- dplyr
- tidyr
- stringr
- ggplot2
- caret
- corrplot
- gam
- Cubist
- zoo
- mgcv
- bst
- sm
- ggpubr
- e1071
- standardize
- ggstatsplot
- kableExtra
- MASS
- pls
- forecast
- nnet
- rpart.plot
- na.tools
- survival
- gtsummary
- magick
- seasonal
- tidyquant
- plyr
- datasets
- lavaan
- eqs2lavaan

D2: List Of Models in the Caret Package

Model	Method Value	Type	Libraries	Tuning Parameters
AdaBoost Classification Trees	adaboost	Classification	fastAdaboost	nIter, method
AdaBoost.M1	AdaBoost.M1	Classification	adabag, plyr	mfinal, maxdepth, coeflearn
Adaptive Mixture Discriminant Analysis	amdai	Classification	adaptDA	model
Adaptive-Network-Based Fuzzy Inference System	ANFIS	Regression	frbs	num.labels, max.iter
Adjacent Categories Probability Model for Ordinal Data	vglmAdjCat	Classification	VGAM	parallel, link
Bagged AdaBoost	AdaBag	Classification	adabag, plyr	mfinal, maxdepth
Bagged CART	treebag	Classification, Regression	ipred, plyr, e1071	None
Bagged FDA using gCV Pruning	bagFDAGCV	Classification	earth	degree
Bagged Flexible Discriminant Analysis	bagFDA	Classification	earth, mda	degree, nprune
Bagged Logic Regression	logicBag	Classification, Regression	logicFS	nleaves, ntrees
Bagged MARS	bagEarth	Classification, Regression	earth	nprune, degree
Bagged MARS using gCV Pruning	bagEarthGCV	Classification, Regression	earth	degree
Bagged Model	bag	Classification, Regression	caret	vars
Bayesian Additive Regression Trees	bartMachine	Classification, Regression	bartMachine	num_trees, k, alpha, beta, nu
Bayesian Generalized Linear Model	bayesglm	Classification, Regression	arm	None
Bayesian Regularized Neural Networks	brnn	Regression	brnn	neurons
Bayesian Ridge Regression	bridge	Regression	monomvn	None
Bayesian Ridge Regression (Model Averaged)	blassoAveraged	Regression	monomvn	None
Binary Discriminant Analysis	binda	Classification	binda	lambda.freqs
Boosted Classification Trees	ada	Classification	ada, plyr	iter, maxdepth, nu
Boosted Generalized Additive Model	gamboost	Classification, Regression	mboost, plyr, import	mstop, prune
Boosted Generalized Linear Model	glmboost	Classification, Regression	plyr, mboost	mstop, prune
Boosted Linear Model	BstLm	Classification, Regression	bst, plyr	mstop, nu
Boosted Logistic Regression	LogitBoost	Classification	caTools	nIter
Boosted Smoothing Spline	bstSm	Classification, Regression	bst, plyr	mstop, nu
Boosted Tree	blackboost	Classification, Regression	party, mboost, plyr, partykit	mstop, maxdepth
Boosted Tree	bstTree	Classification, Regression	bst, plyr	mstop, maxdepth, nu
C4.5-like Trees	J48	Classification	RWeka	C, M
C5.0	C5.0	Classification	C50, plyr	trials, model, winnow
CART	rpart	Classification, Regression	rpart	cp
CART	rpart1SE	Classification, Regression	rpart	None
CART	rpart2	Classification, Regression	rpart	maxdepth
CART or Ordinal Responses	rpartScore	Classification	rpartScore, plyr	cp, split, prune
CHi-squared Automated Interaction Detection	chaid	Classification	CHAID	alpha2, alpha3, alpha4
Conditional Inference Random Forest	cforest	Classification, Regression	party	mtry
Conditional Inference Tree	ctree	Classification, Regression	party	mincriterion
Conditional Inference Tree	ctree2	Classification, Regression	party	maxdepth, mincriterion
Continuation Ratio Model for Ordinal Data	vglmContRatio	Classification	VGAM	parallel, link
Cost-Sensitive C5.0	C5.0Cost	Classification	C50, plyr	trials, model, winnow, cost
Cost-Sensitive CART	rpartCost	Classification	rpart, plyr	cp, Cost
Cubist	cubist	Regression	Cubist	committees, neighbors
Cumulative Probability Model for Ordinal Data	vglmCumulative	Classification	VGAM	parallel, link
DeepBoost	deepboost	Classification	deepboost	num_iter, tree_depth, beta, lambda, loss_type
Diagonal Discriminant Analysis	dda	Classification	sparsediscrim	model, shrinkage
Distance Weighted Discrimination with Polynomial Kernel	dwdPoly	Classification	kerndwd	lambda, qval, degree, scale
Distance Weighted Discrimination with Radial Basis Function Kernel	dwdRadial	Classification	kernlab, kerndwd	lambda, qval, sigma
Dynamic Evolving Neural-Fuzzy Inference System	DENFIS	Regression	frbs	Dthr, max.iter

Model	Method Value	Type	Libraries	Tuning Parameters
Elasticnet	enet	Regression	elasticnet	fraction, lambda
Ensembles of Generalized Linear Models	randomGLM	Classification, Regression	randomGLM	maxInteractionOrder
eXtreme Gradient Boosting	xgbDART	Classification, Regression	xgboost, plyr	nrounds, max_depth, eta, gamma, subsample, colsample_bytree, rate_drop, skip_drop, min_child_weight
eXtreme Gradient Boosting	xgbLinear	Classification, Regression	xgboost	nrounds, lambda, alpha, eta
eXtreme Gradient Boosting	xgbTree	Classification, Regression	xgboost, plyr	nrounds, max_depth, eta, gamma, colsample_bytree, min_child_weight, subsample
Extreme Learning Machine	elm	Classification, Regression	elmNN	nhid, actfun
Factor-Based Linear Discriminant Analysis	RFlda	Classification	HiDimDA	q
Flexible Discriminant Analysis	fda	Classification	earth, mda	degree, nprune
Fuzzy Inference Rules by Descent Method	FIR.DM	Regression	frbs	num.labels, max.iter
Fuzzy Rules Using Chi's Method	FRBCS.CHI	Classification	frbs	num.labels, type.mf
Fuzzy Rules Using Genetic Cooperative-Competitive Learning and Pittsburgh	FH.GBML	Classification	frbs	max.num.rule, popu.size, max.gen
Fuzzy Rules Using the Structural Learning Algorithm on Vague Environment	SLAVE	Classification	frbs	num.labels, max.iter, max.gen
Fuzzy Rules via MOGUL	GFS.FR.MOGUL	Regression	frbs	max.gen, max.iter, max.tune
Fuzzy Rules via Thrift	GFS.THRIFT	Regression	frbs	popu.size, num.labels, max.gen
Fuzzy Rules with Weight Factor	FRBCS.W	Classification	frbs	num.labels, type.mf
Gaussian Process	gaussprLinear	Classification, Regression	kernlab	None
Gaussian Process with Polynomial Kernel	gaussprPoly	Classification, Regression	kernlab	degree, scale
Gaussian Process with Radial Basis Function Kernel	gaussprRadial	Classification, Regression	kernlab	sigma
Generalized Additive Model using LOESS	gamLoess	Classification, Regression	gam	span, degree
Generalized Additive Model using Splines	bam	Classification, Regression	mgcv	select, method
Generalized Additive Model using Splines	gam	Classification, Regression	mgcv	select, method
Generalized Additive Model using Splines	gamSpline	Classification, Regression	gam	df
Generalized Linear Model	glm	Classification, Regression		None
Generalized Linear Model with Stepwise Feature Selection	glmStepAIC	Classification, Regression	MASS	None
Generalized Partial Least Squares	gpls	Classification	gpls	K.prov
Genetic Lateral Tuning and Rule Selection of Linguistic Fuzzy Systems	GFS.LT.RS	Regression	frbs	popu.size, num.labels, max.gen
glmnet	glmnet	Classification, Regression	glmnet, Matrix	alpha, lambda
glmnet	glmnet_h2o	Classification, Regression	h2o	alpha, lambda
Gradient Boosting Machines	gbm_h2o	Classification, Regression	h2o	ntrees, max_depth, min_rows, learn_rate, col_sample_rate
Greedy Prototype Selection	protoclass	Classification	proxy, protoclass	eps, Minkowski
Heteroscedastic Discriminant Analysis	hda	Classification	hda	gamma, lambda, newdim
High Dimensional Discriminant Analysis	hdda	Classification	HDclassif	threshold, model
High-Dimensional Regularized Discriminant Analysis	hdrda	Classification	sparsediscrim	gamma, lambda, shrinkage_type
Hybrid Neural Fuzzy Inference System	HYFIS	Regression	frbs	num.labels, max.iter
Independent Component Regression	icr	Regression	fastICA	n.comp
k-Nearest Neighbors	kknn	Classification, Regression	kknn	kmax, distance, kernel
k-Nearest Neighbors	knn	Classification, Regression		k
L2 Regularized Linear Support Vector Machines with Class Weights	svmLinearWeights2	Classification	LiblineaR	cost, Loss, weight
L2 Regularized Support Vector Machine (dual) with Linear Kernel	svmLinear3	Classification, Regression	LiblineaR	cost, Loss
Learning Vector Quantization	lvq	Classification	class	size, k
Least Angle Regression	lars	Regression	lars	fraction
Least Angle Regression	lars2	Regression	lars	step
Least Squares Support Vector Machine	lssvmLinear	Classification	kernlab	tau

Model	Method Value	Type	Libraries	Tuning Parameters
Least Squares Support Vector Machine with Polynomial Kernel	lssvmPoly	Classification	kernlab	degree, scale, tau
Least Squares Support Vector Machine with Radial Basis Function Kernel	lssvmRadial	Classification	kernlab	sigma, tau
Linear Discriminant Analysis	lda	Classification	MASS	None
Linear Discriminant Analysis	lda2	Classification	MASS	dimen
Linear Discriminant Analysis with Stepwise Feature Selection	stepLDA	Classification	klaR, MASS	maxvar, direction
Linear Distance Weighted Discrimination	dwdLinear	Classification	kerndwd	lambda, qval
Linear Regression	lm	Regression		intercept
Linear Regression with Backwards Selection	leapBackward	Regression	leaps	nvmax
Linear Regression with Forward Selection	leapForward	Regression	leaps	nvmax
Linear Regression with Stepwise Selection	leapSeq	Regression	leaps	nvmax
Linear Regression with Stepwise Selection	lmStepAIC	Regression	MASS	None
Linear Support Vector Machines with Class Weights	svmLinearWeights	Classification	e1071	cost, weight
Localized Linear Discriminant Analysis	loclda	Classification	klaR	k
Logic Regression	logreg	Classification, Regression	LogicReg	treesize, ntrees
Logistic Model Trees	LMT	Classification	RWeka	iter
Maximum Uncertainty Linear Discriminant Analysis	Mlda	Classification	HiDimDA	None
Mixture Discriminant Analysis	mda	Classification	mda	subclasses
Model Averaged Naive Bayes Classifier	manb	Classification	bnclassify	smooth, prior
Model Averaged Neural Network	avNNet	Classification, Regression	nnet	size, decay, bag
Model Rules	M5Rules	Regression	RWeka	pruned, smoothed
Model Tree	M5	Regression	RWeka	pruned, smoothed, rules
Monotone Multi-Layer Perceptron Neural Network	monmlp	Classification, Regression	monmlp	hidden1, n.ensemble
Multi-Layer Perceptron	mlp	Classification, Regression	RSNNS	size
Multi-Layer Perceptron	mlpWeightDecay	Classification, Regression	RSNNS	size, decay
Multi-Layer Perceptron, multiple layers	mlpWeightDecayML	Classification, Regression	RSNNS	layer1, layer2, layer3, decay
Multi-Layer Perceptron, with multiple layers	mlpML	Classification, Regression	RSNNS	layer1, layer2, layer3
Multi-Step Adaptive MCP-Net	msaenet	Classification, Regression	msaenet	alphas, nsteps, scale
Multilayer Perceptron Network by Stochastic Gradient Descent	mlpSGD	Classification, Regression	FCNN4R, plyr	size, l2reg, lambda, learn_rate, momentum, gamma, minibatchsz, repeats
Multilayer Perceptron Network with Dropout	mlpKerasDropout	Classification, Regression	keras	size, dropout, batch_size, lr, rho, decay, activation
Multilayer Perceptron Network with Dropout	mlpKerasDropoutCost	Classification	keras	size, dropout, batch_size, lr, rho, decay, cost, activation
Multilayer Perceptron Network with Weight Decay	mlpKerasDecay	Classification, Regression	keras	size, lambda, batch_size, lr, rho, decay, activation
Multilayer Perceptron Network with Weight Decay	mlpKerasDecayCost	Classification	keras	size, lambda, batch_size, lr, rho, decay, cost, activation
Multivariate Adaptive Regression Spline	earth	Classification, Regression	earth	nprune, degree
Multivariate Adaptive Regression Splines	gcvEarth	Classification, Regression	earth	degree
Naive Bayes	naive_bayes	Classification	naivebayes	laplace, usekernel, adjust
Naive Bayes	nb	Classification	klaR	fL, usekernel, adjust
Naive Bayes Classifier	nbDiscrete	Classification	bnclassify	smooth
Naive Bayes Classifier with Attribute Weighting	awnb	Classification	bnclassify	smooth
Nearest Shrunken Centroids	pam	Classification	pamr	threshold
Negative Binomial Generalized Linear Model	glm.nb	Regression	MASS	link
Neural Network	mxnet	Classification, Regression	mxnet	layer1, layer2, layer3, learning.rate, momentum, dropout, activation
Neural Network	mxnetAdam	Classification, Regression	mxnet	layer1, layer2, layer3, dropout, beta1, beta2, learningrate, activation
Neural Network	neuralnet	Regression	neuralnet	layer1, layer2, layer3
Neural Network	nnet	Classification, Regression	nnet	size, decay
Neural Networks with Feature Extraction	pcaNNet	Classification, Regression	nnet	size, decay

Model	Method Value	Type	Libraries	Tuning Parameters
Non-Convex Penalized Quantile Regression	rqnc	Regression	rqPen	lambda, penalty
Non-Informative Model	null	Classification, Regression		None
Non-Negative Least Squares	nnls	Regression	nnls	None
Oblique Random Forest	ORFlog	Classification	obliqueRF	mtry
Oblique Random Forest	ORFpls	Classification	obliqueRF	mtry
Oblique Random Forest	ORFridge	Classification	obliqueRF	mtry
Oblique Random Forest	ORFsvm	Classification	obliqueRF	mtry
Optimal Weighted Nearest Neighbor Classifier	ownn	Classification	snn	K
Ordered Logistic or Probit Regression	polr	Classification	MASS	method
Parallel Random Forest	parRF	Classification, Regression	e1071, randomForest, foreach, import	mtry
partDSA	partDSA	Classification, Regression	partDSA	cut.off.growth, MPD
Partial Least Squares	kernelpls	Classification, Regression	pls	ncomp
Partial Least Squares	pls	Classification, Regression	pls	ncomp
Partial Least Squares	simpls	Classification, Regression	pls	ncomp
Partial Least Squares	widekernelpls	Classification, Regression	pls	ncomp
Partial Least Squares Generalized Linear Models	plsRglm	Classification, Regression	plsRglm	nt, alpha.pvals.expli
Patient Rule Induction Method	PRIM	Classification	supervisedPRIM	peel.alpha, paste.alpha, mass.min
Penalized Discriminant Analysis	pda	Classification	mda	lambda
Penalized Discriminant Analysis	pda2	Classification	mda	df
Penalized Linear Discriminant Analysis	PenalizedLDA	Classification	penalizedLDA, plyr	lambda, K
Penalized Linear Regression	penalized	Regression	penalized	lambda1, lambda2
Penalized Logistic Regression	plr	Classification	stepPlr	lambda, cp
Penalized Multinomial Regression	multinom	Classification	nnet	decay
Penalized Ordinal Regression	ordinalNet	Classification	ordinalNet, plyr	alpha, criteria, link
Polynomial Kernel Regularized Least Squares	krlsPoly	Regression	KRLS	lambda, degree
Principal Component Analysis	pcr	Regression	pls	ncomp
Projection Pursuit Regression	ppr	Regression		nterms
Quadratic Discriminant Analysis	qda	Classification	MASS	None
Quadratic Discriminant Analysis with Stepwise Feature Selection	stepQDA	Classification	klaR, MASS	maxvar, direction
Quantile Random Forest	qrf	Regression	quantregForest	mtry
Quantile Regression Neural Network	qrnn	Regression	qrnn	n.hidden, penalty, bag
Quantile Regression with LASSO penalty	rqlasso	Regression	rqPen	lambda
Radial Basis Function Kernel Regularized Least Squares	krlsRadial	Regression	KRLS, kernlab	lambda, sigma
Radial Basis Function Network	rbf	Classification, Regression	RSNNS	size
Radial Basis Function Network	rbfDDA	Classification, Regression	RSNNS	negativeThreshold
Random Ferns	rFerns	Classification	rFerns	depth
Random Forest	ordinalRF	Classification	e1071, ranger, dplyr, ordinalForest	nsets, ntreeperdiv, ntreefinal
Random Forest	ranger	Classification, Regression	e1071, ranger, dplyr	mtry, splitrule, min.node.size
Random Forest	Rborist	Classification, Regression	Rborist	predFixed, minNode
Random Forest	rf	Classification, Regression	randomForest	mtry
Random Forest by Randomization	extraTrees	Classification, Regression	extraTrees	mtry, numRandomCuts
Random Forest Rule-Based Model	rfRules	Classification, Regression	randomForest, inTrees, plyr	mtry, maxdepth
Regularized Discriminant Analysis	rda	Classification	klaR	gamma, lambda
Regularized Linear Discriminant Analysis	rlda	Classification	sparsediscrim	estimator
Regularized Logistic Regression	regLogistic	Classification	LiblineaR	cost, loss, epsilon
Regularized Random Forest	RRF	Classification, Regression	randomForest, RRF	mtry, coefReg, coefImp
Regularized Random Forest	RRFglobal	Classification, Regression	RRF	mtry, coefReg
Relaxed Lasso	relaxo	Regression	relaxo, plyr	lambda, phi
Relevance Vector Machines with Linear Kernel	rvmLinear	Regression	kernlab	None
Relevance Vector Machines with Polynomial Kernel	rvmPoly	Regression	kernlab	scale, degree
Relevance Vector Machines with Radial Basis Function Kernel	rvmRadial	Regression	kernlab	sigma

Model	Method Value	Type	Libraries	Tuning Parameters
Ridge Regression	ridge	Regression	elasticnet	lambda
Ridge Regression with Variable Selection	foba	Regression	foba	k, lambda
Robust Linear Discriminant Analysis	Linda	Classification	rrcov	None
Robust Linear Model	rlm	Regression	MASS	intercept, psi
Robust Mixture Discriminant Analysis	rmda	Classification	robustDA	K, model
Robust Quadratic Discriminant Analysis	QdaCov	Classification	rrcov	None
Robust Regularized Linear Discriminant Analysis	rrlda	Classification	rrlda	lambda, hp, penalty
Robust SIMCA	RSimca	Classification	rrcovHD	None
ROC-Based Classifier	rocc	Classification	rocc	xgenes
Rotation Forest	rotationForest	Classification	rotationForest	K, L
Rotation Forest	rotationForestCp	Classification	rpart, plyr, rotationForest	K, L, cp
Rule-Based Classifier	JRip	Classification	RWeka	NumOpt, NumFolds, MinWeights
Rule-Based Classifier	PART	Classification	RWeka	threshold, pruned
Self-Organizing Maps	xyf	Classification, Regression	kohonen	xdim, ydim, user.weights, topo
Semi-Naive Structure Learner Wrapper	nbSearch	Classification	bnclassify	k, epsilon, smooth, final_smooth, direction
Shrinkage Discriminant Analysis	sda	Classification	sda	diagonal, lambda
SIMCA	CSimca	Classification	rrcov, rrcovHD	None
Simplified TSK Fuzzy Rules	FS.HGD	Regression	frbs	num.labels, max.iter
Single C5.0 Ruleset	C5.0Rules	Classification	C50	None
Single C5.0 Tree	C5.0Tree	Classification	C50	None
Single Rule Classification	OneR	Classification	RWeka	None
Sparse Distance Weighted Discrimination	sdwd	Classification	sdwd	lambda, lambda2
Sparse Linear Discriminant Analysis	sparseLDA	Classification	sparseLDA	NumVars, lambda
Sparse Mixture Discriminant Analysis	smda	Classification	sparseLDA	NumVars, lambda, R
Sparse Partial Least Squares	spls	Classification, Regression	spls	K, eta, kappa
Spike and Slab Regression	spikeslab	Regression	spikeslab, plyr	vars
Stabilized Linear Discriminant Analysis	slda	Classification	ipred	None
Stabilized Nearest Neighbor Classifier	snn	Classification	snn	lambda
Stacked AutoEncoder Deep Neural Network	dnn	Classification, Regression	deepnet	layer1, layer2, layer3, hidden_dropout, visible_dropout
Stochastic Gradient Boosting	gbm	Classification, Regression	gbm, plyr	n.trees, interaction.depth, shrinkage, n.minobsinnode
Subtractive Clustering and Fuzzy c-Means Rules	SBC	Regression	frbs	r.a, eps.high, eps.low
Supervised Principal Component Analysis	superpc	Regression	superpc	threshold, n.components
Support Vector Machines with Boundrange String Kernel	svmBoundrangeString	Classification, Regression	kernlab	length, C
Support Vector Machines with Class Weights	svmRadialWeights	Classification	kernlab	sigma, C, Weight
Support Vector Machines with Exponential String Kernel	svmExpoString	Classification, Regression	kernlab	lambda, C
Support Vector Machines with Linear Kernel	svmLinear	Classification, Regression	kernlab	C
Support Vector Machines with Linear Kernel	svmLinear2	Classification, Regression	e1071	cost
Support Vector Machines with Polynomial Kernel	svmPoly	Classification, Regression	kernlab	degree, scale, C
Support Vector Machines with Radial Basis Function Kernel	svmRadial	Classification, Regression	kernlab	sigma, C
Support Vector Machines with Radial Basis Function Kernel	svmRadialCost	Classification, Regression	kernlab	C
Support Vector Machines with Radial Basis Function Kernel	svmRadialSigma	Classification, Regression	kernlab	sigma, C
Support Vector Machines with Spectrum String Kernel	svmSpectrumString	Classification, Regression	kernlab	length, C
The Bayesian lasso	blasso	Regression	monomvn	sparsity
The lasso	lasso	Regression	elasticnet	fraction
Tree Augmented Naive Bayes Classifier	tan	Classification	bnclassify	score, smooth
Tree Augmented Naive Bayes Classifier Structure Learner Wrapper	tanSearch	Classification	bnclassify	k, epsilon, smooth, final_smooth, sp
Tree Augmented Naive Bayes Classifier with Attribute Weighting	awtan	Classification	bnclassify	score, smooth
Tree Models from Genetic Algorithms	evtree	Classification, Regression	evtree	alpha
Tree-Based Ensembles	nodeHarvest	Classification, Regression	nodeHarvest	maxinter, mode
Variational Bayesian Multinomial Probit Regression	vbmpRadial	Classification	vbmp	estimateTheta

Model	Method Value	Type	Libraries	Tuning Parameters
Wang and Mendel Fuzzy Rules	WM	Regression	frbs	num.labels, type.mf
Weighted Subspace Random Forest	wsrf	Classification	wsrf	mtry

Table D0-1 - Caret Models - Source: Adapted from [113]